DOCUMENT RESUME

ED 236 156 TM 830 267

AUTHOR Messick, Samuel; And Others
TITLE National Assessment of Educational Progress Reconsidered: A New Design for a New Era.
INSTITUTION National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO NAEP-83-1
PUB DATE Mar 83
CONTRACT 400-82-0018
NOTE 101p.
AVAILABLE FROM National Assessment of Educational Progress, Box 2923, Princeton, NJ 08541 ($5.00).
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Data Analysis; Data Collection; *Educational Assessment; Elementary Secondary Education; *Federal Programs; Latent Trait Theory; Methods; Policy Formation; Program Content; *Program Descriptions; *Program Design; Research Needs; Sampling
IDENTIFIERS Balanced Incomplete Block Spiralling; *National Assessment of Educational Progress

ABSTRACT

This report presents the conceptual framework and major features of the new design for the National Assessment of Educational Progress (NAEP) as conducted by Educational Testing Service beginning July 1983. It comprises three major chapters. The first chapter reviews the social and environmental changes that demand reconsideration of NAEP. The new design was formulated to address concerns focusing on performance standards, school effectiveness questions, and broad human resource issues, thereby improving NAEP's relevance to educational policy and practice. The second chapter discusses technical innovations now possible with proven modern techniques that greatly enhance the power and value of the collected data. Sampling by grade as well as by age permits estimates of performance and trends to be reported by both age and grade, thereby allowing direct links to state and local assessments, school practices, and educational policies. The third chapter illustrates ways the new design addresses multiple policy questions, communication with multiple audiences in an effective fashion, linkages to other data sources, enhancement and extension of NAEP services, and engagement of the public on the important educational issue of performance standards. Primary type of information provided by the report: Program Description (Operating Policies); Procedures (Conceptual). (PN)

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS RECONSIDERED:

A New Design For A New Era

Samuel Messick
Albert Beaton
Frederic Lord

NAEP REPORT 83-1

naep
The Nation's Report Card

Reading
Math
Science

National Assessment of Educational Progress

ETS

This report was supported by the National Institute of Education under Contract No. 400-82-0018 and Educational Testing Service. The points of view or opinions in this report are those of the authors and do not necessarily represent the position of the National Institute of Education.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS RECONSIDERED:

A New Design For A New Era

Samuel Messick
Albert Beaton
Frederic Lord

in collaboration with

JOAN BARATZ, RANDY BENNETT, RICHARD DURAN, THOMAS HILTON, PAUL HOLLAND, ANN JUNGEBLUT, ARCHIE LAPOINTE, DONALD ROCK, HOWARD WAINER
EDUCATIONAL TESTING SERVICE • PRINCETON, NJ 08541

AND

MORRIS H. HANSEN
WESTAT • WASHINGTON, DC

March 1983

NAEP
The Nation's Report Card

Reading Math Science

National Assessment of Educational Progress

ETS

Preface

This report presents the conceptual framework and major features of the new design for the National Assessment of Educational Progress (NAEP) as conducted by Educational Testing Service (ETS) beginning July 1983.

The new design is *comprehensive* in that it entails procedural changes in sampling, objectives setting, exercise development, data collection, analysis, dissemination, and user services. It is *inclusive* in that the Assessment is extended to previously excluded or inadequately represented populations—in particular, to functionally-handicapped and limited-English speaking students as well as to out-of-school 17-year olds and adults. It is *innovative* in that modern psychometric methology is applied to move the Assessment beyond the level of discrete exercises or arbitrary exercise composites to the level of measurement of performance dimensions. It is *protective* of continuity in that statistical links are forged to past methods and data to maintain and enhance the examination of trends. It is *practitioner-oriented* in that performance data are systematically tied to background and program variables relevant to educational policy and practice. And, it is *aggressive* in its involvement of user groups, educational constituencies, societal stake holders, and the general public to amplify NAEP's impact not only on the conduct of education but on the pluralistic standards and goals of education.

The report comprises three major chapters covering in turn the reasons for the new design, the nature and power of the new design, and the implications and payoff of the new design. The first chapter reviews the strengths and weaknesses of the original assessment design and its responsiveness to the political realities of its time. When social and environmental changes that demand reconsideration of NAEP today are examined, it becomes clear that current national concerns focus on performance standards, school effectiveness questions, and broad human resource issues. The new design was formulated to address these concerns using National Assessment data, thereby improving NAEP's relevance to educational policy and practice.

The second chapter discusses technical innovations now possible with proven modern techniques that greatly enhance the

power and value of the data collected. Through the use of a balanced incomplete block (BIB) spiralling variant of matrix sampling, exciting new analyses are feasible because the data are no longer booklet-bound. Covariances may now be computed among all exercises in a subject area, so that

- composties of exercises can be appraised empirically for coherence and construct validity;
- the dimensional structure of each subject area can be determined analytically as reflected in student performance consistencies;
- item response theory (IRT) scaling can be applied to unidimensional sets of exercises regardless of what booklet they appear in;
- IRT scales can be developed having common meaning across exercises, population subgroups, age levels, and time periods;
- more powerful trend analyses can be undertaken by means of these common scales;
- performance scales can be correlated with background, attitudinal, and program variables to address a rich variety of educational and policy issues; and,
- public use data tapes can be made much more useful because secondary analyses are also no longer booklet-bound.

In addition, groups previously excluded from the Assessment (the limited-English speaking and functionally handicapped) are studied more intensively. Sampling is refined to provide better representation of Hispanic students in terms of their major cultural subgroups (Puerto Rican, Cuban, and Mexican American) and to permit systematic reporting of Hispanic results separately. Sampling by grade as well as by age permits estimates of performance and trends to be reported by both age and grade, thereby allowing direct links to state and local assessments, school practices, and educational policies, which are all typically grade-based. Samples of adults and out-of-school 17-year olds are reintroduced into the Assessment by cost-effective means that also link the exercise performance levels of these groups to labor-force participation data and employment trends.

The third chapter illustrates the ways in which the new design facilitates the addressing of multiple policy questions, communication with multiple audiences in effective fashion, linkages to other data sources, enhancement and extension of Assessment

services, and engagement of the public on the important educational issue of performance standards.

The preparation of this report was partially supported by the National Institute of Education (NIE) under contract No. 400-82-0018, which called for the development of "cost-efficient, imaginative alternative designs to conduct a National Assessment of Educational Progress." It was also included as the lead section on Proposed Design in the successful ETS proposal for "The Conduct of the National Assessment of Educational Progress," in response to NIE Grant Announcement No. PA-82-0001. Now, in order to make the rationale and plans for the redesigned NAEP widely available to a variety of interested publics, the report has become the first release in the new series of NAEP publications under the ETS grant.

Samuel Messick
Princeton, New Jersey
March, 1983

Contents

Contents *(Cont'd)*

Contents *(Cont'd)*

I.
The Original Assessment Design and Changing Assessment Needs

The original design of the National Assessment of Educational Progress (NAEP) was brilliantly responsive to the political constraints of the time. Established in the 1960s to assess the condition and progress of education in the country, the original NAEP design attempted to take due account of the existing political and social realities that were likely to jeopardize its successful implementation. Prominent among these concerns was the recognition that an expanded federal role in education, coming at a time of limited state capacity, represented a serious threat to state and local education agencies. Of prime importance was the feeling that the sanctity of local control of education might be perceived to be undermined by a nationally imposed assessment effort if it conveyed overtones of national curriculum and national testing.

The Politics of Assessment and Its Legacies

In light of such concerns, the original NAEP architects developed a sampling plan insuring that accurate results could not readily be reported at the state or district level. They espoused matrix sampling procedures insuring that no individual would take more than a small sample of diverse exercises or items, so there would be no tests or test scores in the traditional sense and certainly no test scores for any individuals. They capitalized on the strengths of matrix sampling to insure comprehensive coverage in depth within subject matter and in breadth across subject matters, thereby generating sets of objectives and exercises that reflected salient features of most extant curricula but were too extensive to be incorporated in practice in any single curriculum, national or otherwise. They insisted on analysis and reporting at the exercise level, so that the focus would be not on curriculum units or knowledge and skill domains, but on specific learning outcomes whose nature and importance could be directly judged by laymen

and professionals alike. As a final example, the assessment was organized in terms of age levels rather than grade levels, which—while having a number of important points in its favor—has the consequence of severing NAEP results from the major way in which schools are organized, state and local assessments are reported, and educational policies are formulated. Thus, since the original NAEP design by deliberate plan made it difficult if not impossible to link assessment results to state or district programs or to grade-related practices in the schools, educators were less threatened and political feasibility was assured. However, the very design features that were advantageous from a political standpoint also carried the heavy cost of attenuating the usefulness of the assessment results for affecting educational practice.

The Problem of Defensible Interpretations

The main problem with the original assessment design is one of meaning and interpretability of the findings. The intended benefits of exercise-level reporting were simply not realized—namely, that the specific learning outcome embodied in a discrete exercise readily conveyed its own criterion-referenced standard and that a direct link could be easily perceived between the exercise and the educational objective it represented. On the one hand, discrete exercises may often be interpreted to reflect multiple objectives and, on the other hand, it is a rare educational objective of any importance that can be fully captured in a single instance of behavior. Rather, educational objectives refer to consistencies in student performance that cut across classes of behavior (Cronbach, 1971).

This limitation of strict exercise-level reporting of percent-correct on each exercise was eventually addressed by NAEP by also reporting average percent correct on aggregations of exercises presumed to reflect the same dimension or objective. But these aggregations were determined on the basis of educators' judgments and may or may not be supported empirically in terms of student performance consistencies on the exercises judgmentally aggregated. What is needed is not only a means of justifying judgmental exercise aggregations in terms of student performance consistencies, but of empirically determining the aggregations of exercises that best reflect *existing* performance consistencies of educational import. In either case, since the aggregations are interpre-

ted in terms of performance constructs (such as reading comprehension and computational skill), evidence must be accrued for their construct validity and for linking them to educational objectives or sets of objectives as well as to domains of knowledge and skill within subject-matter areas.

The critical requirement for establishing interpretable and defensible aggregations of exercises is to develop a capability for estimating correlations or covariances among exercises as well as between dimensions of exercises and other variables. This capability would permit an empirical evaluation of the coherence and construct validity of the judgmental or nominal exercise categories interpreted in past assessments as "reading comprehension," "science knowledge," and so forth. More importantly, it would permit an evaluation of empirically-grounded exercise categories at different levels of generality, including the possibility of higher-order skills that might cut across content or subject-matter domains. For example, one could appraise the empirical viability not only of exercise categories tightly tied to the behavioral language of task performance, such as "adding two-digit numbers," but also of performance constructs of increasing generality, such as computational accuracy, number facility, and higher-order skills of quantitative reasoning and problem solving. It would also be possible to assess the extent to which higher-order skills such as problem solving and critical judgment cut across subject-matter fields.

By analyzing and reporting assessment results only in terms of specific exercises and unverified judgmental or nominal exercise categories, the relation of trends to more useful indices of achievement is obscured. But by analyzing and reporting empirically-grounded performance consistencies that are interpretable in terms of educationally meaningful dimensions of knowledge and skill and that can be related to other variables of background, attitude, school, and program, the practical and policy implications of the results may be more directly addressed.

The Problem of Comparability

To realize these benefits, however, we need some means of assuring comparability of meaning of performance across exercises within performance dimensions and, of prime importance, comparability across different time periods. Since many factors can

affect percent success on a given exercise, the measurement of change in terms of single exercises is inherently difficult to interpret. Nor do differences in average percent correct across sets of exercises provide satisfactory indices for assessing change. A key problem is that the relationships between percentages and quantitative variables such as those descriptive of background or program characteristics are typically nonlinear, so interpretations of the meaning and sources of percentage change are often either misleading or abstruse. This difficulty may be overcome, however, by employing a scaling model such as Item Response Theory (Lord, 1980a) that transforms percent correct to a logit scale $(\log \frac{p}{1-p})$ to define latent continua which are typically linearly related to other quantitative variables.

An important outcome of this item response theory (IRT) scaling is that exercises are characterized by invariant scale parameters that are directly comparable across exercises on the same latent dimension, whether at the same or different points in time. This enormously simplifies the measurement and interpretation of changes and trends over time. However, to protect and maintain the capability for trend analysis over past as well as future data, the procedural changes entailed in covariance estimation and IRT scaling should be introduced in a way that forges technically viable links to past data.

Although these and other design features are recommended and examined in detail in the body of this report, we are concerned not only with improving the meaning and interpretability of the assessment results but also with enhancing their utilization in affecting educational policy and practice. As a consequence, we will address not only the redesign of data collection, analysis, and reporting procedures but also the redesign of other NAEP activities and functions bearing on objectives setting, dissemination, and knowledge utilization.

Before presenting our recommendations for redesign, however, we will first address the reasons why we think such innovations are feasible in the present political and social context by examining major changes that have occurred in this regard since the 1960s. Then, to insure that our redesigned NAEP will be responsive to current policy issues and flexible enough to respond to changing policy issues, we will next assay the major classes of policy questions that dominate the current educational scene as well as those looming large on the horizon. We are particularly

concerned about those kinds of policy questions that NAEP should be in a position to address but cannot be effectively handled in its present mode of implementation. Next comes the main section of the report which presents the recommendations for redesign in detail and provides the rationale for resolving the major design issues.

Finally, the closing section of the report reviews how the new design improves the meaning and interpretability of assessment results and trends, illustrates its capability for timely response to current and new policy questions and its flexibility for addressing a variety of such questions, and recommends ways of enhancing NAEP's educational impact. The stress in connection with this latter point is on the development of linkages—primarily between NAEP exercises and those used in large or longitudinal research data bases, in statewide assessments, and in commercially published educational tests widely employed in both state and local assessments. By these means the results of research, state, and local studies may be viewed in national perspective and the quality and comparability of assessment at all levels thereby enhanced. Other linkages to be developed are those between the objective setting and standard setting processes and their attendant connections to exercise specifications, performance outcomes, and progress toward the attainment of standards.

Factors Shaping NAEP in the 1980s

The context of education policymaking in the 1980s is significantly different from that of the late 1960s when NAEP was initiated. This section examines the current environment and discusses the policy issues NAEP should be able to address. Of particular importance in understanding the issues and factors presently shaping NAEP are (1) the changed federal role, (2) an increased state capacity for problem solving, (3) an erosion of educational credibility, and (4) the reduction of financial resources. Taken together, along with growing and pervasive pressures for educational accountability, these forces create new demands that must be accommodated if NAEP is to be a useful policy tool in the future.

The Changed Federal Role

Prior to the 1960s the federal government's involvement in education was modest, confined almost exclusively to assisting states with activities they had already adopted. When NAEP was developed, however, the legacy of President Johnson's "Great Society" was in full sway and the federal role had undergone a significant and fundamental change from that of assisting state or local governments to accomplish their own objectives to that of using federal money to accomplish a national purpose (Sundquist & Davis, 1969).

The Elementary and Secondary Education Act of 1965 (ESEA), with its emphasis on disadvantaged children and a focus on building state capacity, was a dramatic and ambitious effort to enlist local and state education agencies in meeting national objectives. Moreover, it served as the centerpiece for a continuing series of measures to extend federal concern to other previously excluded groups: migrants, native Americans, the limited-English speaking, and the handicapped. This new activist thrust of the federal government was the result of two critical assumptions concerning state and local education agencies (SEAS AND LEAS): first, that they either did not know how, or did not fully accept the responsibility, to adequately teach disadvantaged children; and second, that an infusion of knowledge and federal resources could improve the quality of elementary and secondary education.

This expanded federal role represented a threat to many state and local officials in that it not only changed the traditional stance of the federal government in education, but in some instances it conflicted with state and local practices. Distrust was great in both camps: federal officials often felt state and local education personnel were not interested in, or capable of, dealing with federal concerns; state and local administrators feared the imposition of federal regulations and a national curriculum on what had been their time-honored bailiwick of "local control." Passage of the Civil Rights Act of 1964 and subsequent enforcement of school desegregation guidelines under Title VI stoked these fears of federal encroachment.

It was in this environment of tension and distrust that NAEP was designed and implemented. Originally the central question before the developers of NAEP was how to collect representative national data on educational competence while assuring state and local administrators that no federal standard would be imposed nor in-

vidious comparisons made among states or districts. NAEP was merely to be a barometer of the nation as a whole. Usefulness to state and local officials was not a primary consideration.

The 1980s represent a different political environment. The concept of a "New Federalism," with its emphasis on state and local capability for problem solving, hopes to capitalize on the achievements of the past fifteen years of activist federal involvement while attempting to deal with any problems such a federal role created. Not surprisingly, the fifteen year record of federal activism produced both positive and negative effects. Most positive was the adoption of many national objectives and the upgrading of state capacity. Aid for compensatory education is now a feature of 24 state-aid laws (Silverstein et al., 1977). Bilingual education and education for the handicapped have also seen parallel development, with the states in some cases taking the lead and the federal government left to imitate (Wilken & Porter, 1977; Moore, Walker, & Holland, 1982). The negative element of past federal policy on the one hand is a growth in paperwork burden and, on the other hand, the development of statutes and guidelines which, when imposed on the diverse state political cultures, sometimes have impeded rather than enhanced national objectives (Hill & Kimbrough, 1981).

States today may be no less afraid of national standards and curricula, nor should they be, but they appear to be much more open to the use of national comparative information about educational achievement that could help them set their own standards. Although on occasion there have been isolated calls for a "national standard"—for example, by Admiral Rickover during the 1978 hearings on reauthorization of ESEA—such proposals increasingly are viewed as "straw men" and have consistently been opposed by federal education officials on the grounds that setting standards is clearly a state responsibility. The central question now before the directors of NAEP is how to conduct a national assessment that will be directly relevant to state and local policymakers as well as serve as a creditable national indicator of educational competence for the general public.

State Capacity for Problem Solving

When NAEP was being planned, there was a prevalent stereotype of the "backward SEA." In his 1965 testimony urging support for

Title V of ESEA, Commissioner of Education Francis Keppel detailed the weaknesses of state departments, pointing to their lack of staff, inability to monitor and coordinate programs, and general absence of planning activities (Bailey & Mosher, 1968). Since that time there has been considerable upgrading of state department personnel and functions (Murphy, 1973). Virtually all federal elementary and secondary education legislation contains funds for some state department activities—from monitoring and evaluating programs, planning needs assessments and coordinating staff development to increasing equity in school finance formulas. As McDonnell and McLaughlin (1982) point out: "Even those agencies with the fewest resources are able to do more than they could fifteen years ago, and most SEAs are capable of providing significantly more services to local districts." This increased capability is not merely the result of the infusion of ESEA Title V dollars and other federal monies, but also results from state responses to the public cries for accountability and for demands that the educational system "do something" in the wake of bad publicity regarding student performance (McLaughlin, 1981).

Today, state departments of education play a major role in local school improvement efforts (Odden & Dougherty, 1982). They need a wide variety of information on school effectiveness and the relationship of achievement to such factors as school organization, staff training, competency requirements and the like. NAEP should be able to contribute relevant data and analyses to help meet these widespread information needs.

Educational Credibility

When NAEP began, there was some concern about how well the states were serving particular groups, such as the poor and racial or ethnic minorities, as well as serving particular national manpower needs (we were just recovering from the Sputnik shock). But overall there was a belief that the nation's public schools were sturdy, productive institutions. In fact, it was the confidence in schools and their mission that caused the planners of the Great Society to enlist education as the principal soldier in the War Against Poverty (Gardner Presidential Task Force of 1964).

In the 1960s, indeed even into the early 1970s, as the Gallup Annual Education Polls indicate, Americans generally felt their schools were doing a good job (Phi Delta Kappan, 1978). The ma-

jority gave their local schools good grades and believed schools were better than when they themselves had attended. Today the confidence is severely eroded, however, and the majority no longer believes schools are as effective as they had been in the past.

Several factors, some common to institutions in general and others specifically related to education, have contributed to this credibility gap. The disillusionment in the late 1960s and early 1970s with America's involvement in Vietnam coupled with the Watergate revelations of the Nixon Administration served to undermine confidence in many of our traditional institutions—from the Presidency to the military to business to education. But other developments—the SAT score decline, violence and vandalism in the schools, and accounts of illiterate high school graduates—created new demands for accountability. Consumers of the "products" of the education system began to sound the alarm.

The College Board (1977) announced the creation of a Blue Ribbon Panel to investigate the SAT score decline; the Senate held hearings to determine the extent and effect of violence and vandalism in the schools (Bayh, 1977); Pentagon officials argued in Congressional testimony against a volunteer army, citing the lack of preparation of high school youth; businessmen complained about the need to train workers to compensate for the inadequate basic skills of high school graduates; and finally, even students themselves have brought a few malpractice suits against the system for failing to educate them (Baratz & Hartle, 1978). Tales of the educational insufficiencies of young people are commonplace in the media and the cries for relevance, so prevalent in the 1960s and early 1970s, have been replaced by demands for rigor (Fiske, 1981).

One result of the concern about quality was the call for standards. In the early 1970s some states had initiated statewide assessments to monitor general education achievement within their states. In the mid-1970s—with the hue and cry over poor performance of graduates, grade inflation, and social promotion—many states began imposing minimum competency standards on students (and in the late 1970s some states began competency testing for teachers). Within a few years, over two-thirds of the states had minimum competency requirements and virtually every state now has a statewide assessment or minimum competency testing program (Baratz, 1980). "Seat diplomas" were

replaced by specific course requirements and demonstrated competencies for graduation. In the 1980s NAEP should not only assist education agencies to assure high quality assessment programs, but should also facilitate the linking of information now available at the state and local levels with NAEP data. By this means, questions concerning school practices, curricula, progress of particular student groups and the like could be more fully addressed, and assessment results would be more useful to education administrators, classroom teachers, and the taxpaying public.

Fiscal Pressure

NAEP was conceived in the "salad days" of the 1960s when the economy was expanding, enrollments were growing, schools enjoyed the full support of their communities, and federal dollars were increasing. Today the situation is dramatically different.

Since the mid-1970s, there has been a marked increase in the defeat of local school budgets. Even more significant has been the "Proposition 13" phenomenon of the late 1970s—measures limiting taxes and expenditures that severely curtail money available for schools. In addition, along with a general decline in enrollment, there is also a noticeable but modest drift in some regions toward private education (R. L. Smith, 1982) and a lively debate regarding vouchers, tax credits, and other incentives to support private schooling. The state purse almost everywhere is in "ill health" when compared to a decade ago (Shulins, 1982). Tax revenues are not keeping up with inflation. As Adams (1982) observed, four factors are generally responsible for this deteriorating condition: "(1) significant efforts by states to reduce tax burdens, (2) changes in federal individual and corporate income tax structure, (3) a severe recession beginning in 1981, and (4) major cutbacks in federal aid to states and localities."

Demographic changes—declining enrollments, shifts from the cities, increasing numbers of older citizens—have also affected state funding for education. Educators, now more than ever before, find themselves competing with other interests for their share of the public purse. When political competition is coupled with tight dollars in state and local governments, the pressure on educators increases. Meeting the expanding responsibilities of the education system and providing quality education with declining real resources is the major challenge facing state and

local education agencies in the 1980s. NAEP should provide information to state and local officials that is relevant to the effectiveness of various school improvement strategies, information that is not only useful in planning but also addresses state-specific needs.

For all of these reasons, we feel that innovations to improve the interpretability, policy relevance, and utility of NAEP are not only feasible in the current political and social climate, but just about mandatory.

Policy Issues NAEP Should be Able to Address

It seems clear that NAEP must now serve a wide audience with diverse needs. Criticism of NAEP in the past has underscored its failure to be responsive to policy needs (Wirtz & Lapointe, 1982; Milrod, 1980; Wiley, 1981; Sebring & Boruch, 1982). What are some of the issues that NAEP should focus on as it reorganizes to meet the challenges of the eighties?

Among the variety of pressing issues, three general policy areas stand out which should be addressed by NAEP because they require reliable data on student competencies and achievement: student competencies as they relate to *national concerns*; student achievement and attitudes as they relate to *human resource needs*; and, student achievement as it relates to *school effectiveness*. In addressing these issues NAEP must not only be able to provide a national overview, but must also be relevant to state and local concerns—not for the purpose of needless comparisons among states or school districts but to assist individual states and localities in meeting their goals and objectives.

National Concerns

Since NAEP's inception, the federal government has designed and implemented education policies to provide equal educational opportunity to all citizens and to assure that young adults would be able to contribute to society in terms of both productivity and participation in the democratic process. The government clearly

understands that an educated populace is a fundamental requirement for the nation's political and economic well-being. A major responsibility of NAEP should be to provide information for governmental and educational policymakers on the effects of their efforts and to act as an "early warning system" of potential problems.

At a minimum, NAEP data should be relevant to the following kinds of questions:

- Are today's students learning the skills necessary for productive functioning in America in the 1980s? The 1990s? The year 2000?
- Are today's youth developing the flexibility to reorganize their skills in response to occupational and societal change?
- Are students in urban, suburban, and rural schools all being adequately prepared?
- Are public and private school children equally well prepared?
- Do children have access to programs preparing them to deal with the computer age?
- Are minority and disadvantaged youngsters being so prepared?
- Do minority and disadvantaged students in desegregated learning environments perform better than those educated in segregated settings?
- What types of programs or allocations of resources seem to make a difference for disdvantaged and minority students?
- Are children from limited-English speaking homes being provided the necessary skills?
- Do students who have received special services under federal or state programs perform better than similar children who have not had access to those programs?
- Are students developing cultural commitment and appreciation, whether in arts and humanities or in science and technology, or both?
- Do students leave formal education with a positive attitude toward continued learning so essential in our rapidly changing environment?
- Do students leave formal education with positive attitudes toward productive work?

Human Resource Issues

The federal government is concerned with the flow of human resources to assure a work force competent to function in an advanced technology society and the necessary military personnel to protect American interests. Planning for human resource deployment is a complex process that requires reliable information on young people's competencies, training, and attitudes.

In the past we have vacillated between feast and famine in critical personnel areas. In the late 1950s, with Sputnik's launching, we were acutely aware of our need to develop more scientists and engineers. By the late 1960s, however, the market was glutted and engineers and physicists were seeking new careers. Today, once again we find ourselves undersupplied in the science and technology fields, with dim prospects for the future if students do not have a chance to be trained in science and to learn about career opportunities. NAEP should assist governmental and educational policy planners by contributing information on the following kinds of questions:

What are the competencies of students in math and science and what are their attitudes toward these fields?

What kinds of training do students receive?

What are the career goals of high school students?

What are the attitudes of today's youth toward the military? toward business?

To what degree do students with access to science and high technology curricula choose careers in science more than those with no such experiences?

Are we preparing youth to meet the human resource needs in the health sciences? the humanities? teaching?

Are vocational/occupational programs equipping students with the skills they need to function in the work place?

The answers to these questions are of value to business planners, to parents, and to students themselves as well as to educators and government agencies.

School Effectiveness

School administrators are faced with rising costs and multiple demands on limited resources. They must choose among a host of competing interests. Achievement data, to be most useful, should be tied to other information to guide policymakers in deciding how they might best organize their programs and disperse their funds. Although achievement is influenced by many factors—some school related, others beyond the school's control—test data are one measure of the effectiveness of schools. Holding other variables constant, what factors within the purview of school administrators appear most likely to contribute to increased achievement? How can NAEP assist state and local policymakers to improve schooling?

If NAEP is conceived not merely as a social indicator, but as a tool to identify problems and suggest areas of potentially productive research concerning educational progress, NAEP should attempt to provide data that address the following kinds of policy issues:

- Do students in programs requiring minimum competencies and/or graduation test requirements seem to achieve better than other students?
- How do pupil/teacher ratios appear to relate to achievement?
- Do students with preschool and/or kindergarten experiences seem to perform better than those without such programs?
- How do particular curricular approaches relate to student achievement in reading? writing? math?
- What are the relationships of the length of the school year and/or the availability of summer programs to school achievement?
- What are the relationships of in-service training programs, teacher turnover rates, and teacher competency requirements to student performance?
- What types of programs or allocations of resources seem to make a difference in improving school effectiveness?

Although for a number of reasons to be discussed later NAEP is not an appropriate research vehicle to address all of these questions systematically or in depth, timely analyses of the achievement data in relation to relevant background and program vari-

ables should suggest provisional interpretations and promising leads that merit further research attention or special NAEP probe studies.

Implications for Redesign

Henry Acland (1980) succinctly defined the major functions of NAEP: to provide an information base for federal policymakers, to establish a data base for research, to keep track of performance levels, and to help state and local education agencies. NAEP, as originally designed, cannot meet all the demands presently thrust upon it. In order for the assessment to be most useful, it will be necessary to alter some of its practices. The following sections propose ways in which NAEP should be redesigned to address policy issues of the type we have identified here as important to current educational practice. To do this we must attack issues of statistical inference, sampling efficiency, age and grade sampling, timely data collection, covariance estimation, construct validity, dimensional analysis and scaling, trend analysis, correlations with background and program variables, and "causal" analysis.

II.
A New Assessment Design Responsive to Changing Assessment Needs

The proposed redesign of NAEP builds solidly on the original design —but with important modifications, extensions, and innovative additions:

The new design retains the cyclical scheduling of subject-area data collection—*but* (1) changes to a planned schedule of biennial assessment, (2) introduces the assessment of reading into every biennial wave so as to increase the timeliness of information in this basic area as well as to calibrate different cohorts at each age level, and (3) establishes coverage of four subject-matter fields as a minimum target for each assessment wave. The off years are available for focussed studies of special problems or special populations—such as assessing the educational competencies, and in succeeding years the educational progress, of functionally handicapped or limited-English-speaking students. Special assessment probes in areas as yet not covered, such as computer literacy or foreign languages or global awareness, could be conducted either in off years or in connection with a regular assessment wave. In time, NAEP might capitalize on the field presence entailed by special studies during off years to move the assessment of reading and perhaps mathematics to an annual schedule.

The new design retains the current deeply stratified three-stage sampling plan—*but* introduces important additions at the third stage of randomly sampling students within schools so as (1) to effect sizable sampling efficiencies (through the application of a powerful variant of matrix sampling called balanced incomplete block, or BIB, spiralling), (2) to document more fully the characteristics of students presently excluded from the sample as not validly testable by current NAEP procedures, and (3) to undertake sampling by grade level as well as by age. For the second assessment wave in 1985-86, when it would be possible to influence the other stages of the sampling plan, steps would be taken to attain better representation of Hispanic students in terms of their major cultural subgroups (Puerto Rican, Cuban, Mexican American)

and to undertake systematic reporting of the educational progress of Hispanics separately.

The new design retains matrix sampling procedures—*but* as modified in the form of BIB spiralling so as (1) to reduce school clustering effects and thereby sampling errors as well as to produce increased information with a given sample size, (2) to permit IRT scaling of exercises across booklets for objectives and performance dimensions spanning the subject-matter area as well as for those spanning different age levels, and (3) to estimate covariances among exercises. The ability to estimate covariances among exercises within a subject area means that the cohesiveness of judgmental exercise categories can be empirically evaluated, performance categories or dimensions can be empirically determined by methods of factor analysis and cluster analysis, and unidimensionality assumptions of IRT scaling can be empirically appraised. Once exercises are successfully scaled by IRT procedures, pupil proficiency estimates can be related to background, attitudinal, and program variables for the same pupils so that external correlates, and thus the construct validity, of exercise dimensions can be appraised. In the second assessment wave, spiralling of exercises *across* subject-matter areas will permit knowledge and skill dimensions in one area to be empirically related to those in another. Such spiralling will also allow assessment of the degree to which higher-order skills such as inferential reasoning or decision making cut across subject areas.

The new design retains the capacity for comprehensive coverage of subject matter attained through matrix sampling—*but* capitalizes on the structural nature of response consistencies in exercise performance, as appraised or revealed by covariance analysis and IRT scaling, to achieve not only more meaningful or interpretable measurement but more efficient measurement. Thus, basic performance objectives in a field may be effectively measured by structured sets of exercises smaller than those currently used. This would leave more opportunity for the development and use of innovative exercises and for the assessment of higher-order subject-matter skills such as organization, integration, and strategic planning—as for example, in science. These measurement efficiencies will also serve to reduce the number of exercises needed for effective coverage of subject matter in any one assessment wave, thereby yielding important cost efficiencies.

The new design retains the capability for analysis and reporting

at the level of single exercises as well as aggregations of exercises —*but* adds, by means of covariance analysis and IRT scaling, the critical capacity both (1) to construct and evaluate aggregations of exercises in psychometrically responsible fashion and (2) to report the performance of different population subgroups on scales having a common meaning across subgroups, age levels, and time periods. The use of common scales linked across age levels and across time periods enormously simplifies analyses of changes and trends over time while simultaneously yielding more powerful results and straightforward interpretations. Moreover, since both exercises and population subgroups are placed on the same scale, results may be interpreted and reported in either criterion-referenced terms, norm-referenced terms, or both conjointly.

Finally, the new design adds the important capacity to correlate knowledge and skill dimensions with each other as well as with attitudes, interests, background characteristics, and both school and program descriptors, thereby making possible a variety of structural and "causal" or path analyses.

This capsule summary of the critical features of the proposed redesign of NAEP will now be systematically expanded so that measurement, analysis, and cost-effectiveness issues may be addressed in detail.

Data Collection Design Features

The fundamental weaknesses of NAEP are not in the technical quality of its output, which is generally high, but in the limitations of its design and its adherence to procedures of questionable cost-benefit. These weaknesses should be addressed as directly and immediately as possible with due concern for links to past data but not so much concern for past history that the need for change is downplayed or postponed.

Data Collection Schedule

One of the major reasons that NAEP has not become a truly useful indicator of educational progress is that assorted assessment

cycles of three to nine years, which have been characteristic of NAEP in the past, are too infrequent and sporadic either to keep pace with educational change or to keep the public's attention. Worse still, the schedule of subject-matter assessment does not systematically track the student cohorts as they move through the age levels used in sampling and reporting, so that cohort differences are confounded with educational change.

With respect to cohort differences, if a given subject area were assessed in four-year cycles—that is, with three years intervening between assessments of that area then the current sample of 17-year olds assessed in mathematics, for example, would be from the same student cohort as the sample of 13-year olds assessed in math four years earlier and as the sample of 9-year olds assessed in math eight years earlier. Similarly, the current sample of 13-year olds would be from the same student cohort as the sample of 9-year olds assessed four years earlier. By thus matching the assessment intervals to the number of years intervening between the age levels sampled, cohort differences in a given subject area are essentially controlled and interpretations of trend analyses are simplified.

To rectify these problems of timeliness and cohort matching in a cost-conscious way, the proposed redesign entails a planned schedule of NAEP data collection every other year, with reading being assessed biennially. The other two basic areas of mathematics and writing are assessed in alternate waves in four-year cycles, as is science and possibly literature. Because of legal requirements and prior commitments, it is proposed that reading, writing, and citizenship/social studies be assessed in the first year of the redesign (the 15th year of NAEP, 1983-84), but that thereafter four subject areas be covered in each wave so as to shorten the assessment cycle for the remaining learning areas. This proposed assessment schedule is summarized in Table 1 for the first five years of the redesign.

The biennial assessment of reading heightens the pace with which at least one important barometer of national educational progress can be brought before the public and the educational community. In a simple variant of this design, the two basic areas of reading and mathematics would be assessed biennially, which might be possible without sacrificing timely coverage of other areas once the measurement efficiencies discussed below are realized.

The biennial assessment of at least one subject area such as

Table 1

Assessment Schedule for Subject Areas

Assessment Year	Subject Areas			
15th 1983-84	Reading	Writing	Citizenship/ Social Studies	—
16th 1984-85	**Special Studies**			
17th 1985-86	Reading	Math	Science	Area A e.g., Career and Occupational Development
18th 1986-87	**Special Studies**			
19th 1987-88	Reading	Writing	Area B e.g., Literature	Area C e.g., Music/Art

reading also provides an important technical benefit. Although the 13-year old and 17-year old samples collected in assessment year 19 are from the same student cohort as the 9-year old and 13-year old samples, respectively, collected four years earlier in year 15 (which is also true of year 21 samples versus year 17 samples), year 17 represents a different cohort from year 15. Thus, successive waves represent different student cohorts while alternate waves, being spaced at intervals matching the differences in age levels, represent the same student cohort. However, with the assessment of reading common to successive waves, cohort differences can be appraised and calibrated, as it were, and trend interpretations modified accordingly.

The assessment schedule given in Table 1 applies to the three major samples used in NAEP—9-year olds, 13-year olds, and in-school 17-year olds. Although it is important to return to the practice of sampling out-of-school 17-year olds and adults, it is recommended that more cost-effective means be employed for accomplishing this, such as the use of the Current Population Survey of the Bureau of the Census, as discussed in the subsequent section on sampling.

The proposed plan attempts to offset the increased cost of covering four subject areas per wave by deliberately scheduling off years with no data collection every other year. These off years are to be devoted to intensified exercise development, data analysis, report writing, and dissemination. They are also available for

special studies financed through additional resources from a variety of sources. A number of such special studies are briefly described in a later section. Special assessment probes in new subject areas could also be conducted during these off years, again with additional financial resources. But with the capability for correlating across subject areas discussed below, there are advantages to coordinating special probes with the assessment of potentially related or mutually facilitative fields. For example, from the standpoint of illuminating connections and transfer across fields, it would be advantageous to schedule a special probe for computer literacy in year 17 (or 21) when mathematics and science are assessed.

Additional cost-effectiveness further buttressing the feasibility of the proposed schedule derives from the measurement and sampling efficiencies discussed in the later section on spiralling. Since sample size is the major determinant of data collection costs and since the number of exercises answerable in a fixed amount of time drives the number of booklets which in turn drives sample size, improvements in measurement efficiency permitting effective subject coverage with fewer exercises has important cost consequences, as would the negotiation of increased time per student for exercise administration.

Sampling

The proposed redesign retains the current deeply stratified three-stage sampling plan as modified to meet some new purposes in addition to the old. The first stage of sampling entails classifying the primary sampling units or PSUs into strata defined by geographic region and community type. The PSUs are typically counties, but small counties are aggregated so that no PSU has fewer than an estimated 1500 youths at each assessment age. For each age level, the second stage entails enumerating, stratifying, and selecting schools, both public and private, within each PSU selected at the first stage. The third stage involves randomly selecting students within a school for participation in NAEP. For a typical assessment session, from 16 to 25 students of the same age—either 9-, 13-, or 17-year olds—are assembled to respond to the exercises in a particular booklet.

Originally, samples of 17-year old dropouts and early graduates, as well as of adults 26 to 35 years of age, were located in their

homes where one or more assessment booklets were administered. Recently, however, limited budgets have led to less frequent assessment of the adult group as well as the out-of-school 17-year olds, the latter loss being much more serious because of the biases entailed in estimating 17-year old performance from in-school samples alone.

The three sampling stages, with certain exceptions, are acceptable for the proposed NAEP redesign. Some minor procedural modifications are needed at the third stage to accommodate (1) the BIB spiralling variant of matrix sampling, (2) grade-level as well as age-level sampling, and (3) the fuller documentation of the numbers and types of students excluded from the sample as not validly testable by present NAEP means. A modification is needed in the sampling of PSUs and of schools in order to improve the representation of Hispanic cultural subgroups and to permit the systematic reporting of Hispanic performance separately. Sampling of students by grade level, documentation of functionally-handicapped students excludable from the NAEP sample as untestable, and representation of Hispanic cultural subgroups are discussed next in turn; BIB spiralling is treated in detail in the succeeding section.

Sampling by grade as well as by age. The restriction of NAEP to age-level sampling and reporting makes it difficult if not impossible to link national assessment results to school practices, state and local assessments, and educational policies, most of which are typically tied to grade level. This is one of the main reasons that NAEP results are less directly useful than they might be for educational purposes. Accordingly, even though the meaning of grade level varies in different parts of the country depending on the age at which children are admitted to school and on the advancement and retention policies of local school systems, it seems imperative that grade-level sampling and reporting be incorporated into NAEP but not at the expense of eliminating age sampling.

There are also important reasons for sampling by age, not the least of which is that age has a common meaning across geographical regions and school practices. Another critical reason for not relying on grade sampling alone is that *many disadvantaged youth are overage for their grade placement*, which would seriously distort the meaning of average grade-level performance and seriously compromise the interpretation of grade trends as indica-

tions of educational "progress." Taken together, these arguments imply that NAEP sampling and reporting should be by both age and grade.

The addition of grade sampling is not a minor embellishment to age sampling but, rather, a distinctly different though coordinate perspective for characterizing educational achievement and change. According to figures from a recent report of the Bureau of the Census, only about 70 percent of 9-year old students are in grade 4, which is their modal grade, and a roughly similar percentage of students in grade 4 are nine-year olds, which is the modal age in that grade. Similar percentages hold for 13-year olds and grade 8 while somewhat lower percentages obtain for 17-year olds and grade 12. Hence, age and grade sampling and their associated analyses provide critical counterpoint to each other in disentangling the import of performance levels and trends. In addition, following the lead of Truman Kelley (1940), special analyses of the "ridge" of students of modal age who are in their modal grade might provide useful norms for many comparative purposes, although they might also be simplistic for other interpretive purposes.

Documenting sample exclusions. Although NAEP is meant to be a barometer or report card on the national condition of education, past implementations have excluded significant populations of students from data collection in particular, limited-English speaking and functionally-disabled pupils. The exclusion of these populations has significant implications for NAEP both because of their size and the resources invested in their members' education. While the exclusion of these populations limits the generality of the NAEP report card, such exclusion is understandable because many practical and theoretical issues exist in the assessment of both handicapped and non-English proficient students.

In the past, NAEP has dealt with these issues by directing local school personnel to exclude students falling within three gross categories: limited-English speaking, functionally disabled, and educable mentally retarded (Research Triangle Institute, 1979). Criteria for determining membership in these categories has been left primarily to the judgment of the local school districts. Data collected on these excluded cases appears to have been limited solely to the number of pupils falling within each broad category. These categorizations obviously provide precious little information on exactly *who* is being omitted from the NAEP program.

Knowing who is being excluded from NAEP is critical for at least

two reasons. First, without such information, it is difficult to know precisely whom the NAEP report card does and does not represent. For example, NAEP data on Hispanics may not be representative of Hispanic youth as a whole due to the exclusion of non-English proficient students from data collection. Second, if the NAEP barometer is truly to represent the national condition of education, we must eventually find meaningful and practical ways to assess currently excluded populations. Detailed description of these populations is a necessary first step in developing workable assessment strategies for them. Since much of the needed information is contained in student records that can be consulted by school officials and trained data collectors as part of the process of identifying students to be excluded from the assessment, its systematic collection would be facilitated by the development of a form for characterizing excluded students along a number of important background dimensions.

The proposed form would include such pupil and program descriptors as age, sex, ethnicity, languages of the home and frequency of use, current program (duration, setting, percent time mainstreamed, related services, pupil/teacher ratio, primary goal areas, languages used in instruction, percent of instruction in English), years of previous special or language instruction, type and severity of handicapping condition, and specific reason for exclusion from NAEP.

Within the proposed NAEP redesign, three major uses of these types of data are envisioned. First, such data will provide a meaningful characterization of students excluded from NAEP samples and hence from generalizations about results. Second, this characterization will be compared with other characterizations of handicapped and limited-English speaking students formulated from existing data bases (e.g., those generated through periodic surveys conducted by the Office for Civil Rights, the National Center for Education Statistics, and the Annual Child Count of PL 94-142). This comparison should suggest the extent to which special segments of the handicapped or non-English proficient populations—such as the learning disabled—are being served by NAEP. Finally, the data collected will be employed as the basis for a proposed strategy for assessing traditionally excluded groups in future years, a strategy discussed at greater length in the later section on Special Studies.

Sampling Hispanic students. Given the increasing size of the Hispanic population in the United States and the distinctive edu-

cational problems of Hispanics related to bilingual and bicultural background, Hispanic results should not be averaged together with those of other groups but rather should be analyzed and reported separately, as has been done to some degree in recent NAEP reports. In doing this, however, it would be important to attain representative coverage of the major Hispanic cultural subgroups—Puerto Rican, Cuban, and Mexican American—because of differences in their social and migrational histories that have implications for their educational progress. Since these groups are differently distributed throughout the country, this implies some modification of the sampling plan. This change in the sampling procedures would not be initiated before the second assessment wave in the NAEP redesign (1985-86), when it would first be possible to influence the various stages of the sampling design.

In addition, there remain two other sampling issues that warrant further discussion, each entailing possibly cost-effective compromises with current or former procedures. One involves a strategy for administering NAEP exercises to adult samples and possibly to out-of-school 17-year olds. The other involves the enlistment of cooperating schools for repeated participation in NAEP.

Sampling adults. Since competent adult functioning in society is an ultimate goal of educational progress, it is important for NAEP to return to the practice of sampling adults. Furthermore, since estimates of 17-year old performance based only on in-school samples are inevitably biased, it is important to include out-of-school samples as well. It is proposed that cost-effective means for accomplishing this be seriously investigated, such as the use of the Current Population Survey of the Bureau of the Census.

Every month the Census Bureau surveys 70,000 households to ask a variety of questions, using a continuously rotating sample. All contacts are made during the same week of each month by 1500 trained part-time permanent employees who visit or telephone each of the 70,000 households. Each household is used eight times during a 16-month period—households are in the sample during four consecutive months in one year, out eight consecutive months, and back in the sample the same four calendar months the next year. Each month there is a 75 percent overlap with the previous month's sample. The sample is highly stratified using Census data—PSUs are generated at the county level and about 250 areas are in the sample with certainty, some 160 of which are metropolitan areas. In addition, approximately

375 other PSUs are sampled, about 40 of which are metropolitan areas. The samples are updated monthly using construction and building-permit records or, where those are not available, actual physical inventories of housing units are listed. Each October the school-enrollment study is conducted, and NCES is considering becoming a regular co-sponsor of that effort. Non-Census government agencies may participate in the Current Population Survey with supplementary inquiries, but are limited to fifteen minutes per interview in a particular month.

Preliminary inquiries indicate that NAEP, as a government-sponsored program, is eligible to participate and that administration of subject-matter exercises is considered to be feasible, although they might be restricted to the concluding segment of interview sessions. Since in-home administration of NAEP exercises would require special training of the interviewers, the expected lead time might exceed the current estimate of six to seven months. Moreover, since the collection of labor-force participation information for the Department of Labor is a major part of this service, it might be possible to relate educational achievement measures on samples of adults and out-of-school 17-year olds obtained by this means directly to indicators of labor-force participation and employment trends.

Repeated school participation. When independent samples of schools are drawn in successive assessment waves, school-to-school differences in average performance level contribute part of the sampling error in the measurement of changes over time. Therefore, sizable reductions in sampling error could be attained if the same schools participated in successive assessment waves. From the standpoint of both sampling efficiency and school contact costs, it would seem ideal to recruit schools to participate in four successive assessment waves, with a fourth rotated out and replaced by a new sample of schools each wave. Realistically, however, this strategy might produce an unacceptably deleterious effect on the cooperation rate. Furthermore, participation of a school in one assessment wave might affect its performance in succeeding waves. Therefore, although this strategy should be seriously investigated, it does not seem highly promising and should be carefully evaluated before proposing its implementation.

A compromise between independent sampling of schools in successive waves and repeated participation of the same schools may prove more feasible and still substantially improve efficien-

cy. This compromise entails rotation of PSUs and schools in the sample so that 50 percent of the PSUs and schools are identical in two successive assessment waves for the same subject area. The advantage of this compromise over the current NAEP approach of independent school sampling is that it should substantially reduce the sampling errors of measures of change over time. That is, schools make important contributions to variance for any given assessment wave and, with independent samples in successive waves, school contributions to the variance of the differences are essentially doubled. With an identical sample of schools in the two waves, these contributions to the variance of change are reduced by a factor of (1-*r*), where *r* is the average within-PSU correlation between the years for the particular exercise or aggregate being estimated for the identical schools. Since r is often as large as .7 or more, worthwhile efficiencies are achieved in estimates of change. A rotating sample with 50 percent of the schools identical from one wave to the next would achieve about half of this benefit.

Unless school cooperation can be retained at substantially the same level under this procedure and unless participation in one wave affects performance in the next only moderately at most, this compromise strategy should not be adopted. However, since a rotation group effect observed in several studies tends to approach a modestly biased but stable level over time, this compromise strategy should be carefully appraised. It might prove feasible in connection with state assessments, in which cooperating states could arrange for school participation in a two-wave rotational plan.

Balanced Incomplete Block (BIB) Spiralling

The theoretical basis for the current method of assigning exercises to respondents by matrix sampling was developed at Educational Testing Service by Frederic Lord (1955, 1962). Matrix sampling, as implemented by NAEP, entails dividing the exercise pool for a given age level into different assessment packages or booklets such that each package contains about as many exercises as a student can answer in the given time period. The packages are discrete in the sense that an exercise that appears in one package does not usually appear in another, although exercises often appear in other packages at a different age level. This method of

matrix sampling is adequate for estimating the proportion of persons in a population who can respond correctly to an exercise. It is not adequate for determining the structure of performance consistencies in a subject area or for estimating levels and trends in composite variables created from exercises in different assessment packages.

Another technique for distributing exercises to respondents is conventional spiralling, which has long been used by Educational Testing Service in its major testing programs. As an example, each Scholastic Aptitude Test (SAT) contains one section that does not contribute to an individual's SAT score but is used instead for introducing new and innovative items and for linking the present test form with past and future forms of the SAT. Although each individual takes only one such variable section, it is possible to administer a number of different sections in a single SAT administration. This is done by spiralling the variable section—that is, test booklets are assembled so that, say, the first booklet has variable section 1, the second booklet has variable section 2, and so forth, until all variable sections have been distributed and then the process is repeated. Since examination booklets are assigned to individuals in the order in which they are seated in the examination room, administration is easy as long as the variable sections all require the same amount of time. Pre-coded answer sheets are inserted in test booklets so that the different sections are distinguishable by scoring machines.

The proposed NAEP redesign entails a modified data collection procedure that combines the advantages of matrix sampling with those of conventional spiralling. This procedure, which is called balanced incomplete block (BIB) spiralling, is an extension of ideas expounded by Knapp (1968). Essentially, it involves developing a balanced incomplete block design such that each exercise is administered the same number of times as it would be in matrix sampling, but in addition each *pair* of exercises is also assessed a prescribed number of times. This means that each exercise will be located in several different packages or booklets, so that many different packages must be printed for an exercise pool of a given size. The BIB spiralling of exercises also implies that many different packages, and thus different sets of exercises, will be administered in a particular assessment session.

BIB spiralling and matrix sampling. An example contrasting ordinary matrix sampling with the BIB spiralling variant of matrix

sampling will illustrate their differences. Consider a reading assessment for age 13 and assume that the assessment pool contains 165 exercises. Assume further that a 13-year old can do 33 reading exercises during the allotted assessment time and that the sampling plan calls for 2,100 13-year olds to take each exercise. Although these assumptions are arbitrary, they are reasonably close to what would be expected during a typical assessment wave. These particular numbers were chosen to simplify the arithmetic below.

The matrix sampling approach as employed by NAEP would divide the exercise pool into five different packages of 33 exercises each. Each different package would be bundled into sets containing as many copies of that package as there are students expected in an assessment session. A selected school would be assigned one or more assessment sessions and would receive one or more different sets of packages accordingly. Following past practice, no school would receive all packages. For each assessment session, a different random sample of students within the school would be selected and scheduled. All students in a given session would receive the same set of exercises because of the current NAEP practice of taped aural presentation and pacing. A sampling and management plan is needed to assure that each set of packages is administered an appropriate number of times within each PSU. The total assessment would include five packages each administered to 2,100 youths or 10,500 students in all.

Next, consider the BIB spiralling approach. First, it is clear that a distinct package cannot be developed for each possible combination of exercises since the number of combinations of 165 exercises taken 33 at a time is astronomical. In the balanced incomplete-block-approach, however, the exercises can be combined into 15 discrete blocks of 11 exercises each, and these blocks of 11 exercises can be permuted such that each pair of blocks occurs together in at least one package. Under this plan, many more different packages would be printed, although the number of students taking each exercise as well as the total number of students assessed would be the same as in the present NAEP matrix sampling plan.

A balanced incomplete block design that fits these specifications is shown in Table 2. The blocks of 11 exercises are numbered from one to fifteen. Each row of the table shows the numerical designation of the blocks that would be contained in a particular package; the left-hand set of columns shows the blocks

Table 2

Balanced Incomplete Block Design for 165 Exercises in 35 Booklets Comprising 33 Exercises Each*

Booklet No.	Simple Block Order			Random Block Order		
1	1	2	3	1	4	11
2	1	4	5	15	9	8
3	1	6	7	14	4	8
4	1	8	9	1	6	14
5	1	10	11	9	12	6
6	1	12	13	11	2	5
7	1	14	15	13	3	2
8	2	4	7	8	5	3
9	2	6	5	14	12	13
10	2	8	11	13	5	6
11	2	10	9	12	3	7
12	2	12	15	14	5	9
13	2	14	13	12	10	8
14	3	4	6	7	9	2
15	3	5	7	5	10	7
16	3	8	10	13	4	10
17	3	9	11	9	10	11
18	3	12	14	8	1	7
19	3	13	15	4	6	7
20	4	8	12	15	12	11
21	4	9	13	15	6	10
22	4	10	14	8	6	2
23	4	11	15	13	9	1
24	5	8	15	3	14	15
25	5	9	12	4	3	9
26	5	10	13	15	5	4
27	5	11	14	3	10	1
28	6	8	14	2	12	4
29	6	9	15	2	10	14
30	6	10	12	13	15	7
31	6	11	13	7	11	14
32	7	8	13	6	3	11
33	7	9	14	1	15	2
34	7	10	15	1	5	12
35	7	11	12	11	8	13

*Each booklet contains 3 blocks of 11 exercises each (33 exercises per booklet). There are 15 blocks of 11 exercises (165 exercises total).

in simple order and the right-hand set shows the same design with the package numbers randomly recoded and the rows and columns randomly permuted. This design would require that 35 different assessment packages be printed, each package containing three blocks of 11 exercises.

Examination of the table indicates that each block of exercises occurs in exactly seven packages and that each pair of blocks occurs in precisely one package. If each package is administered 300 times, then each block of exercises will be presented to 2,100 different students. An exercise in one block will be administered to the same students as an exercise in another block 300 times. The total assessment would include 35 packages times 300 students for each package or 10,500 13-year olds in all, the same number as in the matrix sampling design.

Moreover, BIB spiralling simplifies the administration of assessment sessions. Under the present NAEP application of matrix sampling, care must be taken to distribute the correct packages within PSUs. Consider now that the 35 different packages in the BIB spiralling example are merged in a random sequence and that the same sequence is repeated for all sets of 35 packages. If for a target assessment session of 25 students the packages are assembled in consecutive sets of 26 or 27 packages, then each session will have enough packages for the scheduled students and one or two extra in case of special situations. Under this cycling system, each package will be first in a set an equal number of times and the packages not used at the end of a set will be balanced over all sets. Thus, within a PSU the only consideration is the number of assessment sessions or the total sample size, and the particular packages administered is not a management issue.

It should also be stressed that BIB designs, although not necessarily available for exercise pools of any particular designated size, may be readily developed for a wide array of sizes. Indeed, we have not yet found a reasonably sized pool for which an appropriate design could not be developed in that neighborhood. For example, although there is not a good design for 100 exercises in blocks of ten, there is an excellent design for 99 exercises in blocks of 11. An example for 250 exercises in blocks of ten appears in the later section on IRT scaling. Designs may be of many types: Latin squares, Youden rectangles, lattices, and so forth (Cochran & Cox, 1957). In any event, if no balanced design can be found for a particular case, a slightly less efficient imbalanced design could be used instead.

Although ordinary matrix sampling in this example requires only five different packages while BIB spiralling requires 35, the total number of printed packages or booklets, as well as the total number of printed pages, remains the same. For the extra assembly costs, we have assured that each *pair* of blocks of exercises is administered to a certain number of youths. In this way a complete cross-products matrix of all exercises can be produced, and this matrix can serve a number of important functions—such as ascertaining the interrelationships among objectives or performance dimensions, testing the unidimensionality of the measurement area or subareas for applications of IRT scaling, and delineating the structure of achievement in an area by means of factor analysis and multidimensional scaling. It should be noted, however, that this cross-products matrix is not quite a standard one because its elements are based on different samples; the analytic features of this type of matrix are discussed in a later section on covariance analysis.

It should also be noted that BIB spiralling is statistically more efficient than ordinary matrix sampling for some estimates. By administering more different exercises within a particular school and by administering a particular exercise in more different schools, the school clustering effect is reduced and the BIB sampling design is consequently more efficient. Preliminary calculations, using reasonable assumptions about the cluster effects now common in NAEP results, suggest that BIB spiralling can reduce the number of students necessary to attain a given sampling error by about 20 to 25 percent when compared to ordinary matrix sampling, or reduce the standard errors by 10 to 15 percent when using the same sample size.

In the proposed redesign, BIB spiralling is applied in the first assessment wave only in the assessment of reading. This is because data collection for citizenship/social studies is the completion of an assessment already begun using the original matrix sampling procedure, and the repackaging of the writing exercises as currently constituted is of dubious cost-effectiveness. However, in the next and succeeding assessment waves, BIB spiralling is to be applied in all subject areas. In addition, BIB spiralling will be undertaken *across* subject areas to delineate interconnections between knowledge and skill in one area and that in another as well as to appraise the degree to which higher-order skills cut across areas.

From public use tapes to public USEFUL tapes. Except for simple analyses of average percent correct on aggregations of exercises judged to assess the same objective, current NAEP data based on ordinary matrix sampling is inherently booklet-bound. For judgmental aggregations of exercises that cut across booklets, analyses going much beyond the simple reporting of performance levels face a major roadblock. Even appraisal of the empirical coherence and correlates of such aggregates must be undertaken one booklet at a time, if at all. This is a serious limitation in primary NAEP data analyses, but it is even more debilitating in secondary analyses based on the current public use data tapes.

Each data file on these public use tapes contains the results of one booklet or package of exercises for one age level of one assessment wave. This means that even for simple analyses of average percent correct that entail aggregation across several packages in a subject area, it is necessary to process from 10 to 30 separate data files. Just to locate all of the exercises written for a particular objective or containing a particular type of subject-matter content requires an elaborate use of tables. Worse still, any appraisal of the reliability or generalizability of the exercises representing a specific objective, as well as appraisals of their construct validity vis-à-vis correlations with other objectives or with background variables, must be carried out one booklet at a time on whatever collection of exercises happens to appear there (Anderson, Welch, & Harris, 1982; Hambleton, 1982).

One of the major benefits of BIB spiralling is that both primary and secondary analyses of NAEP data are freed from the booklet bind. Correlations can be computed among all exercises in a subject area. Any aggregation of exercises from whatever combination of booklets can be appraised for reliability or generalizability and correlated with other item aggregations as well as with background variables. Similarly, IRT scaling can be applied to exercises drawn from any set of booklets in the subject area. Secondary analysis is also enormously facilitated by public use data files each of which will now contain all of the exercises in a subject area easily retrievable by objective measured, by type of content, by format, and so forth. In short, BIB spiralling makes it possible to convert public use tapes into public USEFUL tapes.

Trade offs in taped aural administration. The use of BIB spiralling has one serious implication that must be confronted, which is that BIB administration is inconsistent with aural presentation and pacing of exercises using a tape recorder. This is not likely to

seriously affect the reading assessment, which is only paced by tape and not presented aurally. But each of the other areas may be substantially affected. The problem, of course, is that with BIB spiralling the students are assessed on different packages, and aural presentation would result in cacophony at the assessment session, unless expensive equipment such as headphones were employed.

Since taped presentation and pacing would be forgone with BIB spiralling, the cost-benefits of the trade off must be appraised. On the one hand, poor readers, whether from disadvantaged minority groups or not, perform somewhat better with aural as well as printed presentation, while good readers appear not to be unduly distracted on the average—although some good readers are undoubtedly distracted. On the other hand, aural presentation is expensive and requires extra equipment as well as some special skills at the assessment session.

Much more important, aural presentation and pacing of exercises is a procedure common to NAEP but rare indeed in other educational measurement enterprises. No state or local assessments, to our knowledge, have adopted aural presentation procedures and it is doubtful that any will. This renders NAEP procedures and hence NAEP results noncomparable to the mainstream of educational measurement practice. Innovative procedures such as taped presentation are only of marginal value if they cannot reasonably be used in state assessments or other testing programs. Costs aside, the major trade off thus appears to be between improved measurement validity for some students and comparability of results of all students. Our conclusion, therefore, is that aural presentation and pacing of exercises has questionable cost-benefit while BIB spiralling has considerable and multiple cost-benefits.

However, since the same exercise presented by printed page alone will almost certainly have different properties than when presented aurally as well, past NAEP results cannot be expected to be comparable with those obtained in the redesigned NAEP if tape presentation is eliminated. For this reason, statistical links must be established to the past data in each area to maintain the capability for trend analysis. Procedures for establishing these links are discussed in the next section.

In summary, at the cost of increased printing and assembly expenses and the aural presentation of exercises, BIB spiralling simplifies administration, reduces sampling error, and provides the

ability both to determine the dimensionality of a subject area and to develop scales using the most powerful available methodology. It also results in data that can be more usefully organized on public use tapes and more meaningfully described in reports and in the public media.

Statistical Links to Past Data

As has just been emphasized, changes in the method of presenting exercises may affect the probability of a correct response for some students and hence the proportion of correct responses for various groups. Thus, comparisons of proportions or of average percent correct over the time interval spanning the method change could be misleading because method differences are confounded with educational trends during this period. Yet, not changing the method of presentation commits NAEP to a perpetuation of expensive procedures that restrict the comparability and utility of NAEP results and hinder the implementation of powerful innovations like BIB spiralling. The solution is to forge statistical links to the past so as to permit translations from past data based on one method to new data based on another. The capacity to make such translations would effectively maintain the integrity of trend analyses across the method change. This statistical linking, then, is a means of both preserving what has been done in the past and of moving responsibly into new methodology.

Equating samples. The proposed statistical link essentially requires an equating study in which data are collected on some student samples by the past method and on other student samples by the new method during the same assessment wave in each affected subject area. There are three types of data sets at issue in this equating strategy:

Set A contains data from past assessments collected using taped presentation procedures. These are the data whose usefulness for trend analyses we wish to preserve.

Set B contains data from a future assessment wave collected using precisely the same taped presentation methods as in set A. Since the data in sets A and B were collected by the same method but at different times, any differences between them are attributable either to educational change or to sampling error. Since sampling error can be estimated, so can the amount of educational change.

Set C contains data collected in the same assessment wave as set B but using the new methodology. Since the data in sets A and C were collected not only at different times but by different methods, any differences between them are attributable to method differences as well as to educational change and sampling error. More information is needed to disentangle these three components and thereby render the data in set C comparable to that in set A in the estimation of educational change.

The leverage for solving this problem comes from the data in set B. If B and C are based on random samples from the same population, then the differences between them are attributable only to method differences and sampling error. Since they were collected in the same assessment wave, they do not differ in educational change. The data sets B and C can be used to estimate the effect of method, thus disentangling method differences and educational change in comparisons with set A.

Therefore, whenever a substantial change in data collection methodology is introduced, the NAEP redesign entails an equating study to estimate the effect of the method change. Essentially, this involves collecting data by the old method as well as the new for different random samples from the same population of youth. Data need be collected by the old method for set B on only those exercises from set C that are repeated from past assessments. Data collected by the new method for set C should be based on a full-sized sample of students so that sampling error is not increased and so that set C is directly comparable to future data collected by the new method. It is proposed that set B be based on half-sized samples, however, which our calculations indicate should be sufficient for equating purposes.

Equating methods. Composite variables comprising several exercises can be easily and straightforwardly equated using standard methods, such as equipercentile equating, provided that some of the blocks used in spiralling for set C are constrained so that packages in set B can be composed of sets of those blocks. More powerful equating methods using item parameters from IRT scaling are applicable only if the method effect is small. If the method of presentation affects low scorers more than high scorers, then the requirement that the logistic function (relating the probability of a correct response to proficiency or ability) should be the same in both groups would clearly be violated. Equipercentile equating would avoid this anomaly and provide a simple and clear comparison of composite scores obtained by the two methods.

Equating of single exercises is less straightforward and requires some psychometric development. At a minimum, the differences in response proportions due to changing methods of presentation and to sampling error would be described. In addition, a simple nonlinear function fit to the proportions found by the two methods could be used for purposes of translation and adjustment.

This equating-sample approach also has the advantage of protecting against unforeseen problems. If the change in method of presentation radically and massively affects the results, then the equating sample B is available for comparison with the past and continuity is maintained, albeit with a larger sampling error because of the smaller sample size. In this case, a decision would need to be made as to which method to use in the future. If it were decided to continue the former method, then full-sized samples would be collected by that method in future waves. In the more likely case of deciding in favor of the new method—because of cost-effectiveness, analytic power, and comparability of results to other assessment programs—then trend data would be plotted discontinuously. The earlier data would be presented along a time line ending at the point corresponding to the equating sample, and the later data would be plotted along the same time line but beginning with a different value based on set C.

In the proposed NAEP redesign, only reading will be spiralled during the initial wave; writing and citizenship/social studies will be administered by the past matrix sampling procedures with taped presentation and pacing. Since in the past, reading exercises have not been aurally presented by tape but have been paced by tape, an equating sample is proposed to appraise the effects of changing from paced to unpaced administration. Because the method effect of pacing is likely to be small, at least in comparison with the effect of aural presentation, IRT equating may be applicable; hence, it may be possible to represent time trends on a common performance scale spanning past and future reading data. In future assessment waves, spiralling is contemplated for all subject areas, and an equating study will be undertaken for each area as it is introduced.

Analysis Design Features

The introduction of balanced incomplete block spiralling into data collection has profound implications for data analysis. To begin with, it makes possible the computation of covariances among exercises within a subject area and, in future assessment waves, across subject areas as well. In addition, it facilitates the application of IRT scaling to exercises in different packages, thereby yielding scales that span a subject area and, ultimately, scales with a common meaning that span age levels and time periods as well. The integrative properties of IRT scaling in turn have powerful implications for trend analysis and for correlational and "causal" or path analysis of relationships between performance scales and background, attitudinal, and program variables.

Covariance Analysis

Since BIB spiralling assures that each pair of exercises is responded to by a specified number of individuals, a covariance matrix can be computed among all of the exercises in a subject area and, in future assessment waves, between exercises in one area and those in other areas assessed at the same time. In light of this latter capability, it would make sense to select subject areas for a particular assessment wave that are mutually facilitative, like science and mathematics, so that the transfer relationships of knowledge and skill can be appraised.

The availability of covariances among exercises provides a number of immediate benefits. First, it contributes to construct validation (Cronbach, 1971; Messick, 1975, 1980) in that the coherence of exercises designed to measure the same objectives can be empirically evaluated, as can the degree to which an exercise relates to other objectives for which it was not intended. It is possible, for example, that a graph-interpretation problem in social studies is more closely related to mathematics exercises than to other types of social studies exercises. From this discriminant aspect of construct validity, a second benefit of covariances is obtained—namely, undesired method variance can be detected and corrected. Thus, by identifying exercises that assess the same dimensions or objectives regardless of exercise format, the generalizability of interpretations becomes empirically grounded. This

is not to imply that graph interpretation should be excluded from the assessment of social studies, but rather that it should not be combined with social studies exercises that measure a different dimension of knowledge or skill. Contrariwise, it does suggest that the content coverage of graph problems in mathematics could be enriched by inclusion of social studies material. In any event, the decision about what kinds of exercises to include in an assessment must be based *both* on expert judgment about relevance and coverage and on demonstrated response consistencies in student performance (Loevinger, 1957; Messick, 1975, 1980).

A third benefit of covariances is economy of measurement. By empirically grouping sets of exercises that reliably assess a common dimension or objective, composite scores can be used which entail smaller sampling errors. Preliminary calculations indicate that, by going from one exercise to a composite of ten exercises, sampling error is cut roughly in half but that further reductions in sampling error diminish as the number of exercises in the composite increases. With covariances available, item analysis procedures could be used to refine large composites to optimal or cost-effective levels of reliability and sampling efficiency.

In short, covariances provide an empirically-grounded conceptual basis for defining meaningful scales and scores. This will move NAEP from the level of statistical description of performance on single exercises or unverified judgmental aggregations of exercises to the level of measurement.

The structure of educational achievement. Moreover, the entire matrix of intercovariances, or selected submatrices, can be analyzed by such multivariate methods as metric and nonmetric factor analysis and multidimensional scaling to ascertain the dimensional structure of performance in the domain. In this connection, it should be noted that the covariance matrix generated via BIB spiralling differs from the usual covariance matrix in that its elements are based on different random samples of individuals. This means that the overall matrix, because of sampling and measurement errors, may not be consistent with cross-products generated from any single set of real scores. If this is the case, principal components analysis of the covariance matrix would yield at least one dimension having a negative sum of squares, a mathematical inconsistency indicating that the matrix is not appropriately analyzable by standard multivariate methods. However, an effective solution is to estimate a population covariance matrix, which will always be consistent and hence analyzable by

standard methods (B. Wingersky, 1982). Therefore, covariance matrices based on BIB spiralling will be tested for consistency and adjusted accordingly, if necessary, before undertaking dimensional analyses.

Another technical difficulty warrants further comment. In binary response data of the type obtained with exercises scored correct or incorrect, the covariances are distorted by curvilinearities in the relationship between exercise responses and the underlying performance dimension. If the exercises vary only moderately in difficulty level, this problem is handled by using tetrachoric correlations, especially if they are corrected for the effects of guessing (Carroll, 1961). But if the exercises differ widely in difficulty, it may be necessary to use alternative approaches such as nonlinear factor analysis (McDonald, 1983) or methods that attempt to fit the factor model directly to the binary data (Tucker, 1983). Since this problem may be effectively finessed by factor analyzing not item scores but composite scores for small exercise clusters, this approach will be applied as well, using both empirically-verified rational composites of exercises and those derived by homogeneous-cluster keying.

Appropriate factor analyses of covariance matrices among exercises will be employed in the NAEP redesign to ascertain the dimensional structure of each subject area. The performance dimensions isolated will be compared with the objectives specified in exercise construction to identify any commonalities. Those dimensions that cut across the original objectives will be carefully examined to see if process interpretations can be educed suggestive of new, more process-oriented objectives or of higher-order skills. Depending on the outcome, the factor analysis may thus yield dimensional scores for existing objectives as well as scores for unanticipated dimensions that cut across the existing objectives. In any event, the analysis will illuminate the structure of performance in the domain, which should have important implications both for instruction and future measurement.

Group differences in structure. The issue of fairness in measurement impels us to inquire whether performance dimensions have the same meaning and are measured with the same precision in different population groups. This issue of population generalizability will be addressed for separate dimensions by IRT scaling in the next section. Here, we propose the application of confirmatory factor analysis of covariance structures in different groups of the same age to see if the same number of dimensions

emerge in each group and if they are related in the same way (Jöreskog & Sörbom, 1979). This method will be applied to the comparison of male and female groups as well as to black and white groups at each age level and, ultimately, to other groups of special interest.

However, when the data are broken down by sex or by race, for example, the covariances obtained for minority groups from BIB spiralling may be based on small samples, although oversampling may obviate this difficulty in some instances. Hence, in sone circumstances it may first prove necessary to use Bayesian missing-data techniques to adjust the sparse data by capitalizing on prior knowledge of the total covariance matrix and the cells of the group covariance matrix based on sizable samples (Dempster, Laird, & Rubin, 1977).

It is important to ascertain whether the covariance or factor structure in different groups is the same or not because the interpretation of group differences in mean level of performance depends upon it. Indeed, multivariate statistical tests on means assume an invariant covariance structure. Once similarity of the underlying factor pattern in different groups is established, however, the interpretation of mean differences becomes legitimate in the sense that there is supporting empirical evidence that they reflect discrepancies along the same dimensions. If only some of the factors are invariant across groups while others appear to be group specific, then comparisons of group means on the invariant factors would be reasonable. Other differences in factor structure might be less benign, however, in their impact on the interpretation of mean differences. If the same factor structures are found to hold in the different groups, the equality of measurement precision across groups may also be tested by this confirmatory factor model (Jöreskog & Sörbom, 1979; Rock, Werts, & Grandy, 1981).

Age differences in structure. Similarity or difference of covariance structures in different age groups may also be analyzed in the same manner by this confirmatory factor model. Of particular concern in age-group comparison is the possibility of developmental trends not only in mean level of performance but in the degree of differentiation and integration of the skill dimensions at different age levels. There are numerous theories of human development supported by considerable empirical evidence that an individual's cognitive skills and achievements become more differentiated over time (e.g., Kagan & Kogan, 1970; Guilford, 1967). This would in turn be reflected in differences in the factor inter-

correlations among these dimensions at different age levels. Using confirmatory factor analysis, we can address this possibility in each subject area by testing for differences in the factor variances and intercovariances across the three age groups of 9-, 13-, and 17-year olds.

Since we are also concerned about whether age-related differences in factor differentiation occur in the same way for all sex and race groups, similar age-group comparisons will be conducted, if the resulting sample sizes permit, separately for male and female and for black and white groups. Again, any obtained age-group differences in the number and nature of underlying factors will have critical implications for the interpretation of mean differences between the age groups, because that would imply that the same dimensions are not being measured or are not being measured in the same way at different ages.

Scaling by Item Response Theory

Item response theory (IRT) defines the probability of answering an exercise correctly as a mathematical function of ability level or skill. The particular mathematical function most widely used, the logistic function, has one parameter for each individual—namely, ability level—and from one to three parameters characterizing each exercise (Lord, 1980a; Lord & Novick, 1968). The item parameters reflect difficulty level, discriminating power, and likelihood of guessing. The three-parameter model will be emphasized here because the one- and two-parameter versions do not adequately cope with the realities of exercise variation.

IRT methods are appropriate for unidimensional areas or subareas in which the exercises are scored right, wrong, or no response. In the 1983-84 NAEP assessment, reading is the only area for which IRT methods will be fully used, although subareas of citizenship/social studies and possibly multiple-choice writing items will also be analyzed. In subsequent years, IRT scaling will be used for mathematics, science, and other appropriate areas. The possibility of using IRT models for exercises having other scoring formats, such as those scored on a scale from 0 to 10, will also be investigated (e.g., Samejima, 1972, 1973, 1974). The following description of the rationale and procedures for data collection and analysis will typify IRT methods to be used in areas having dichotomously-scored exercises, such as reading and mathematics.

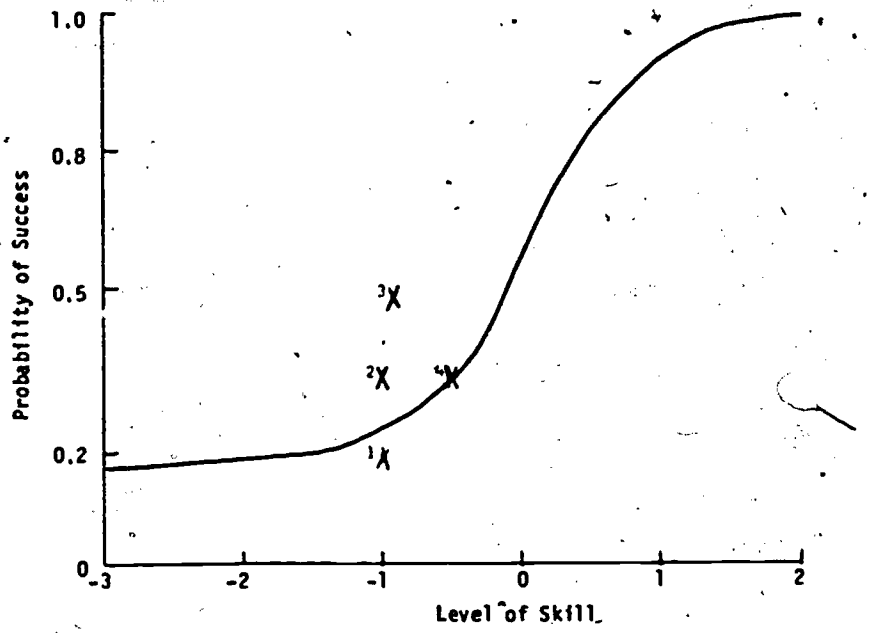
Individual- versus group-based IRT scaling. In the proposed NAEP data analyses, the IRT model to be employed will fit the responses of individuals, not some group mean of individuals. Although IRT models defined at the level of groups, such as schools or demographic subgroups, have been proposed (Bock, Mislevy, & Woodson, 1982), it seems hardly plausible to assume that subgroup mean performance has a true functional relationship to mean level of skill in the subgroup.

From this standpoint, such group IRT models seem fundamentally flawed at a theoretical level, as may be seen from the following example. Figure 1 shows a typical item response function representing the performance of individual respondents on a given exercise. The four crosses mark the mean performance levels on this exercise of four different hypothetical schools or subgroups. The students in the first subgroup or school (lowest cross) all have skill levels that are tightly distributed about −1, and thus about 20 percent of these students will answer the exercise correctly. In the second group, the range of skill is from −2 to 0, and some 35 percent of the students answer correctly. In the third group, the range of skill extends from −3 to 1, and about 50 percent answer correctly. In the fourth group, the range is from −1 to 0, and about 35 percent answer correctly. Although the example is an extreme one, it clearly demonstrates that mean subgroup performance, whether at the level of schools or of demographic categories such as those in the sampling design, cannot be expected to have a true functional relationship to mean level of skill. Thus, such group-based models do not fulfill the fundamental requirement of IRT methodology, which is that the probability of answering correctly be a mathematical function of ability level or skill.

Dimensionality. Since IRT models, whether individual- or group-based, are applicable only to unidimensional sets of exercises, the availability of covariances will be capitalized on to meet this requirement. Factor analyses will be carried out to determine how the exercises in a skill area can be subdivided into subareas that are roughly unidimensional. In the mathematics area, for example, exercises may be classified into the following categories: calculation, story problems, geometry, definitions, measurement. In one approach, a group factor will be extracted for each of these subareas and the residuals examined to see if there are other significant group factors needed to account for the item intercorrelations. The correlation of each group factor with

Figure 1

School Means Plotted in Relation to Exercise Response Function



the general factor for all the exercises will also be computed. We will then decide whether a general factor may be substituted for all or some of the group factors without serious loss.

In this way we will be able to appraise whether all or nearly all reading exercises (or mathematics or science exercises) can be analyzed together in IRT work. If a few exercises do not fit this procedure, they will be removed from the IRT analysis and analyzed by conventional methods such as proportion-correct. If the exercises fall into two or more subareas that cannot be merged, each such subarea will be treated separately for IRT analysis, provided it contains enough items for this purpose.

Assessment. With BIB spiralling of exercises, IRT methods may be applied to exercises appearing in different packages—indeed, if unidimensionality is satisfied, to all of the exercises in a subject area. For example, Table 3 shows a balanced lattice design allocating 25 blocks of different exercises among 30 subgroups of students within a given age group. If there are 12,000 students altogether, then each exercise is taken by 2,400 people. If there are

250 exercises altogether, each block contains ten exercises. Each student answers five blocks or 50 exercises. The existing computer program LOGIST (M.S. Wingersky, 1982; M.S. Wingersky, Barton, & Lord, 1982), which we propose to use to estimate each individual's level of skill or proficiency, is designed to handle sparse data matrices such as Table 3.

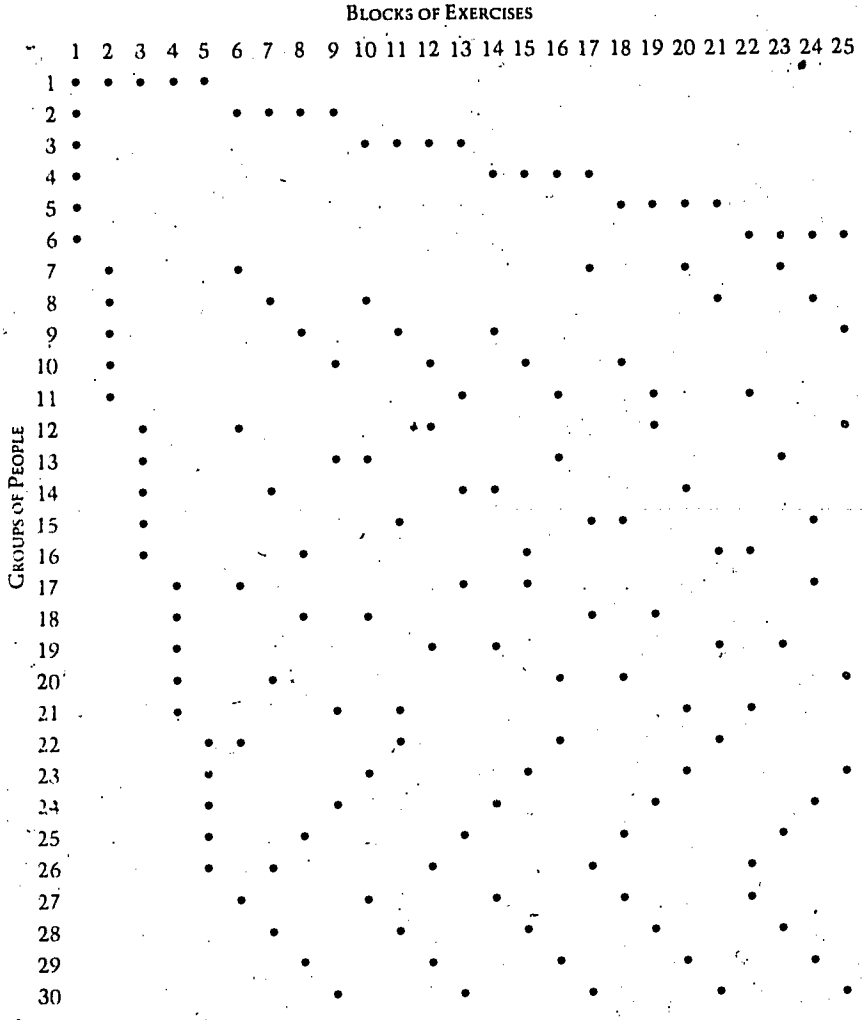
There is, of course, no intention of reporting skill levels for individuals. Rather, the assessment of groups, which is the ultimate purpose of NAEP, will be accomplished by the pooling of individual assessments. This assessment of the individual is given by the maximum likelihood estimate of his or her level of skill under IRT assumptions (Lord, 1980a). In NAEP applications, each individual term in the maximum likelihood equations can be weighted by the sampling weight assigned to the individual in the sampling frame. One efficiency of LOGIST is exemplified by noting that the computer time used is proportional to the amount of data (to 2400 × 250 = 12,000 × 50 = 600,000 responses in the illustrative example), not proportional to both the number of exercises and the number of people simultaneously (not to 250 × 12,000 = 3,000,000).

LOGIST uses a three-parameter logistic model for the data. Its final output consists of one number for each individual that assesses skill level and three numbers that describe each exercise: one for the difficulty of the exercise, another for the extent to which success on the exercise is related to the overall assessment in the area scaled, and a third number representing the proportion of successes on the exercise among very unskilled individuals. This last number, which is often ignored or misused, should not be neglected during NAEP assessment.

The success level for unskilled individuals, denoted for exercise i by c_i, is necessarily nonzero for multiple-choice items, which can be answered correctly by guessing. The usual oversimplifications assume that all $c_i = 0$ (one-parameter or Rasch models and two-parameter models) or that all c_i are equal across exercises. It is also commonly but mistakenly asserted that c_i cannot be accurately estimated. Figure 2 is presented to contradict all these views. It shows c_i estimated by LOGIST from two different data sets for the same exercises. The exercises plotted are all those for which $b_i - 2/a_i > -2$, where b_i is the IRT difficulty parameter and a_i is the discriminating power. It is clear from Figure 2 that the c_i can be reliably estimated for exercises that are discriminating and not too easy.

Table 3

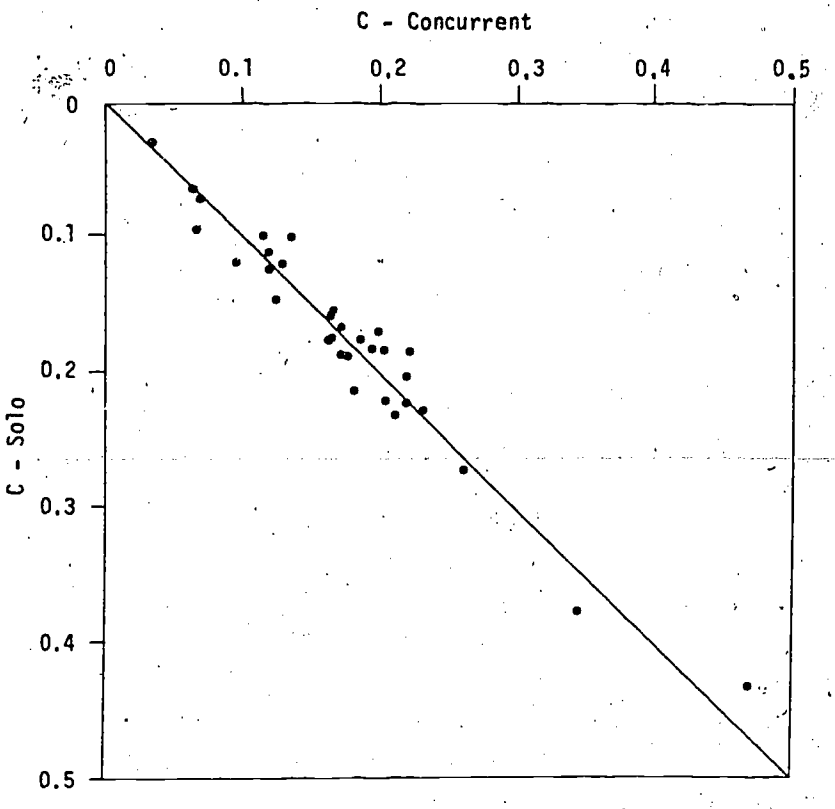
Balanced Lattice Design Allocating Exercises to People



Suppose an educational statistician assumes that all $c_i = .2$ for a large set of NAEP exercises. The data will very likely contradict this assumption. For example, the statistician will later find that one exercise was answered correctly by only 11 percent of all individuals in a certain large socioeconomic subgroup.

Figure 2

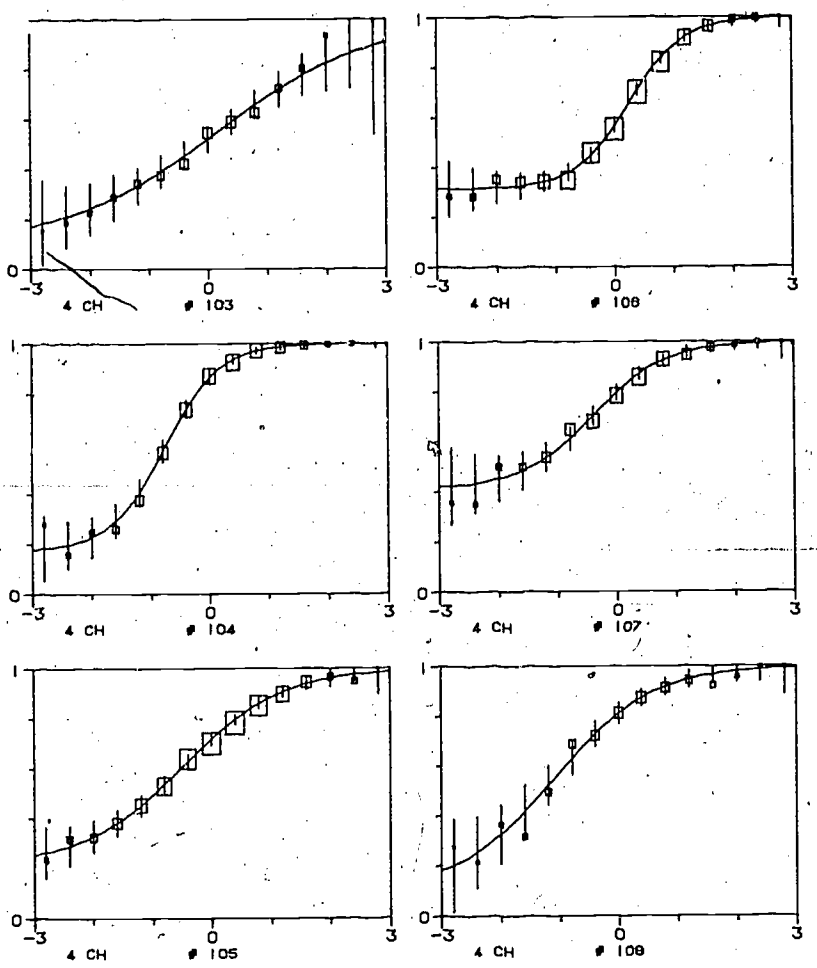
Comparison of Estimated C Values
SAT ESAO4 March '82
Solo vs Concurrent



LOGIST also affords a solution to a problem in the current mode of NAEP reporting. When NAEP reports that 30 percent of individuals in a certain subgroup answered a particular four-choice exercise correctly, it is difficult to interpret this number. If individuals who had no idea of the correct answer guessed at random on the exercise, the 30 percent has a different meaning than if all such individuals either omitted the exercise or indicated they did not know the answer. The recent NAEP practice of reporting average percent correct across exercises judged to represent a particular objective or achievement area simply exacerbates the problem.

Figure 3

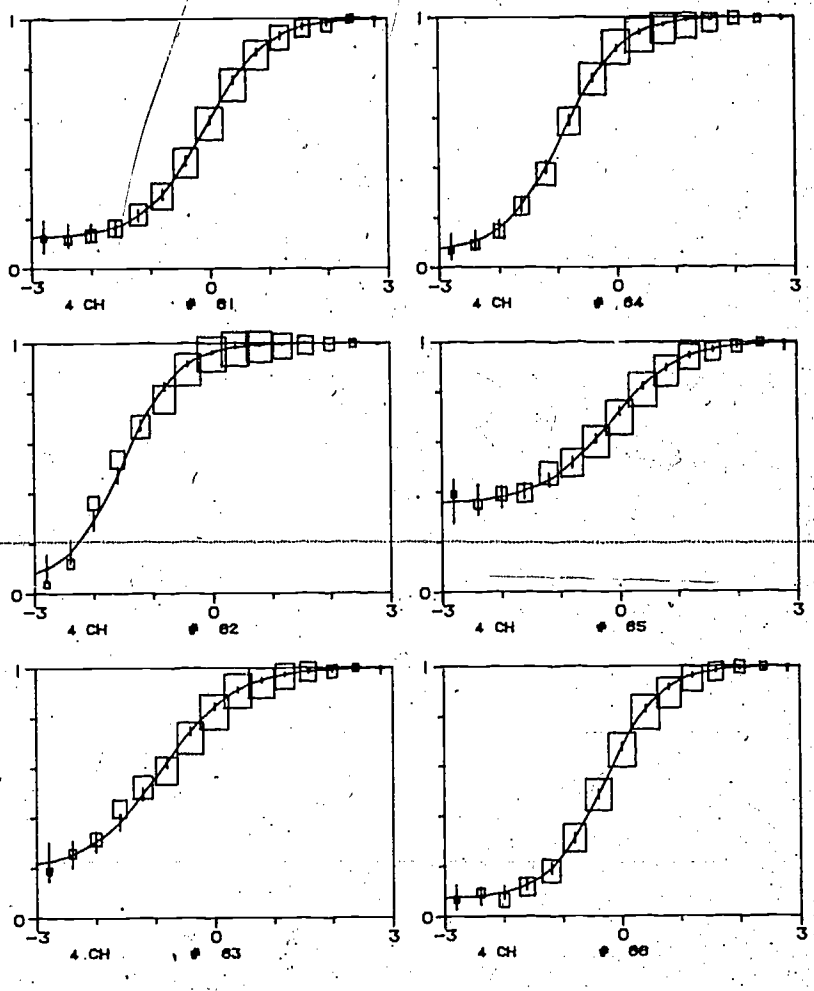
New Jersey Basic Skills Results for
Six Reading Comprehension Exercises



In IRT work, it is seriously incorrect to treat omitted or "do not know" responses the same as wrong responses. It is also incorrect to treat omitted or "do not know" responses as if the corresponding exercises had not been administered. Currently, LOGIST is the only IRT program to our knowledge that treats such data in a reasonably appropriate manner (Lord, 1974).

Figure 4

New Jersey Basic Skills Results for
Six Mathematics Computation Exercises



Checking IRT model fit. Figures 3 and 4 show the type of plot that has been used in all IRT operational work at Educational Testing Service for the past three years in order to provide a visual check on how well the IRT model is able to fit the data. The smooth curves in each figure are estimated response functions for six consecutive four-choice exercises in the New Jersey College

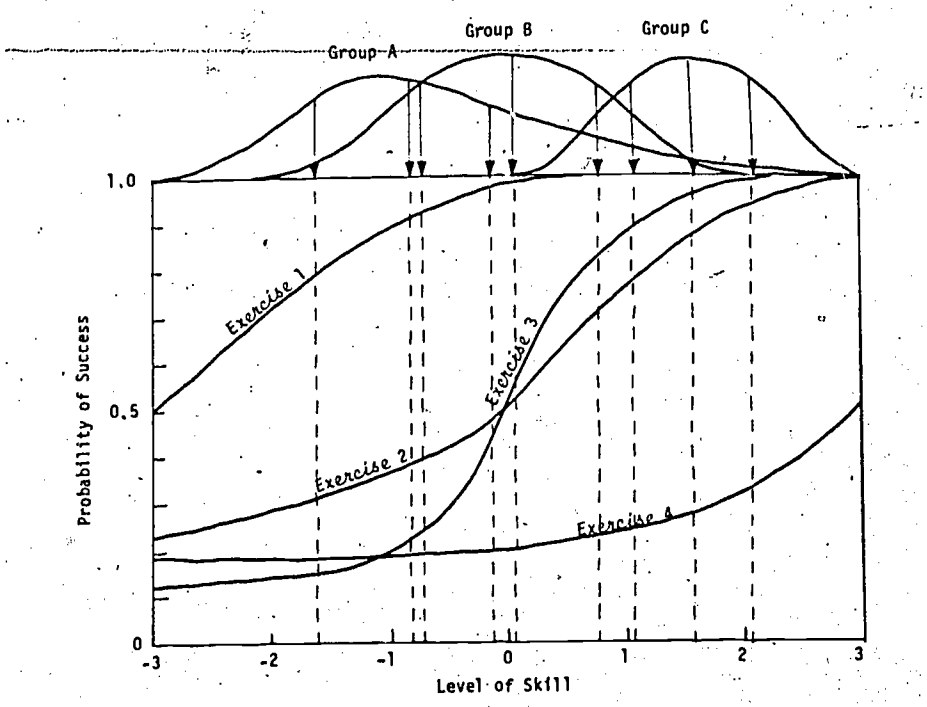
Basic Skills Placement Test. The horizontal axis in each plot shows the skill of the respondent; the vertical axis shows the probability of a correct answer.

Respondents are divided into 15 class intervals according to their estimated level of skill. The area of each plotted rectangle or square is proportional to the number of examinees in the corresponding class interval. The center of the rectangle indicates the observed proportion of respondents in the interval who actually answered the exercise correctly. The vertical line in each interval extends two binomial standard errors above and below the theoretical curve.

Figure 3 presents results for six reading comprehension exercises. The data came from a sparse matrix such as that in Table 3. The number of examinees for these items ranged from 2,400 to 9,600. Figure 4 shows results for six mathematics computation exercises. Each plot represents the results for 21,000 to 24,000 examinees.

Figure 5

Distributions of Skill in Three Subgroups Together with Expected Performance Levels on Various Benchmark Exercises



Examination of these plots convinces us that (1) unskilled examinees have better than zero chance of success, (2) their chance of success varies sharply from exercise to exercise, (3) their chance of success on the more difficult exercises can be accurately estimated, and (4) the slopes of the curves (the discriminating power of the exercise) vary sharply from exercise to exercise.

A chi square comparing theoretical and observed frequencies is also computed for each plot. It is helpful to list the contribution of each class interval to this chi square. Although this procedure, like other available procedures (Hambleton, 1982), does not permit an exact test of statistical significance, it has nevertheless proved helpful in locating ambiguous or other anomalous exercises that clearly do not fit the IRT model. Such exercises can be studied by conventional methods based on proportions of correct answers.

Estimating group performance on a common scale. The main purpose of the IRT analyses is to provide a common scale on which performance can be compared across groups and subgroups, whether tested at the same time or several years apart. IRT allows us to estimate group performance for any group or subgroup, even though all respondents did not take all the exercises in the NAEP pool.

A technical report of results will contain many figures such as Figure 5, showing the distribution of skill in various subgroups together with expected performance levels on various benchmark exercises. The vertical arrows mark the median and the first and third quartiles in the distribution of skill for each specified group. The figure can be read to give the proportion of correct answers on each exercise expected for individuals at each quartile (or at any other point) in each group. It can also be read to give the proportion of individuals in any group who have less than some specified probability of success on any given exercise. More accurate information will also be given in numerical tables.

The actual text of the benchmark exercises will accompany such figures and tables. Note that this provides a *criterion-referenced interpretation* of the meaning of each numerical level of skill: the skill score is interpreted in terms of expected performance on typical, benchmark exercises. *Norm-referenced interpretations* are also provided by such figures and tables by reference to the group distributions.

Appraising item bias. If an exercise has exactly the same item response function in every group assessed, then individuals at any given skill level will have exactly the same probability of getting the exercise correct, regardless of their group membership. This is true even though some groups may have a lower average skill level than other groups. However, if an exercise has a different item response function for one group than for another, then the item is biased in some way.

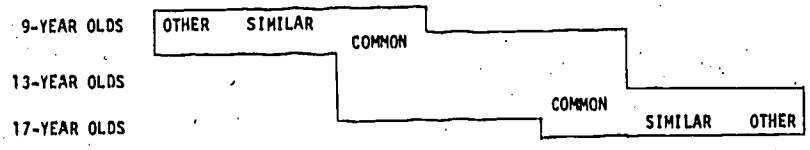
If the item response function for one group is higher than that for another at all levels of skill, then individuals in the first group have a better chance of getting the exercise correct than individuals of equal ability in the second group. A more complicated form of bias occurs if the item response functions for two groups cross, as is often found in practice, because then the exercise is biased in favor of some members of each group but against other members. If item bias is substantial, the exercise should be omitted from the LOGIST run and studied by conventional methods, if at all. These types of bias can be evaluated using IRT methods by estimating item parameters separately for each group and comparing the item response functions across groups (Lord, 1976, 1980a).

Development of a common scale across age levels. Table 3 illustrates the assignment of exercises to individuals in the same age group. Many exercises are given both to 9-year olds and to 13-year olds; many others are given both to 13-year olds and to 17-year olds. This design is indicated in Figure 6. Each row of Figure 6 has a fine structure like that in Table 3.

The exercises in the top and bottom rows of Figure 6 are divided into three categories: (1) those exercises that are common to two age groups, (2) those that are similar in topic and in difficulty to these common exercises, and (3) other exercises. The main LOGIST run discussed above will not be limited to any single age group. Rather, it will include all data in Figure 6 except for the exercises marked "other" for 9- and 17-year olds; all exercises for 13-year olds will be analyzed. This will place all age groups on the same skill scale. After this has been done, each individual's estimated skill level will be held fixed while the parameters describing the "other" exercises are found by a further LOGIST run.

Measuring change across time. Exercises in NAEP reading administrations prior to 1983-84 were administered in printed form, with taped pacing but without the use of taped aural presentation. If the effect of pacing proves minimal, the 1983-84 reading

Figure 6
Assignment of Exercises Within and Across Age Groups



scale can be and will be extended to all exercises administered in past years. This could be done by the method described in the previous section.

A preferable procedure will be to make separate LOGIST runs on data from earlier NAEP administrations. Exercises common to an earlier administration and a later administration will be used to place all earlier results on the same scale as the 1983-84 results. This will be accomplished by the computer program TBLT in current use at ETS (Stocking & Lord, in press). This program finds the linear scale transformation that places two sets of IRT parameters on the same scale in such a way as to minimize a certain sum of squared errors. The quantity minimized is the mean squared difference between number right scores on the common items predicted from the two sets of IRT parameters that are to be placed on the same scale.

By the same method, future groups assessed in reading without taped pacing can be compared on a common reading proficiency scale with groups assessed in 1983-84. Furthermore, if the effect of pacing proves to be minor, these future groups can also be compared on a common reading scale with groups assessed in previous NAEP administrations. Similar comparisons can be made for mathematics and other areas, except that the use of aural tape presentation before 1983 may impair attempted common-scale comparisons extending backwards in time before 1983.

The power of IRT scaling. Among the considerable benefits of IRT scaling for NAEP is the availability, for strictly analytical purposes, of weighted composite scores for each individual on unidimensional aspects of the subject area in which he or she was assessed. This means that performance dimensions in each subject area may be correlated both with each other and with background, attitudinal, and program variables tied to these same students. Furthermore, a variety of subgroups may be defined in terms of these variables—such as bilingual versus monolingual, large school versus small school, Title I participation versus none,

or science interest versus arts interest. The educational performance of these constructed groups may then be compared, if the resulting subgroup sizes are adequate, either simply or with covariance controls for other variables. Moreover, when spiralling occurs across subject areas, the correlational structure of performance scales and their correlates may be addressed both within and across subject fields.

Another benefit of IRT scaling is invariance both of item parameters across respondent groups and of respondents' skill levels across subsets of exercises. This means that each individual's skill level may be estimated from any subset of exercises and that exercises may be added or retired from the assessment at any point without affecting comparability of results. Furthermore, since the skill scales are unbounded, they are not warped by floor and ceiling effects in the way percentages and total scores are, so they tend to be more linearly related to other quantitative variables. These advantages combined with those previously discussed—especially the capacity for both criterion-referenced and norm-referenced interpretations and for linking overlapping sets of exercises to form common scales spanning subject area, population subgroups, age levels, and time periods—make IRT scaling not only ideal for NAEP purposes, but essential.

Analysis of Time Trends

There are a variety of opportunities for studying time trends in the data gathered in the initial wave of the NAEP redesign in combination with data from previous waves of NAEP. The availability of trend data for the subject areas covered is summarized as follows:

Reading:	70-71, 74-75, 79-80, 83-84.
Writing:	69-70, 73-74, 78-79, 83-84.
Citizenship:	69-70, 75-76, 81-82, 83-84.
Social Studies:	71-72, 75-76, 81-82, 83-84.

Thus there are four waves of data for each of the four subject areas to be assessed in 1983-84. The methods of trend analysis discussed below are applicable to time-structured data of this type and hence may be employed with past waves of data in other subject areas or with future waves of data.

Two different levels or types of trend analyses are proposed. The first is at the level of individual exercises, and the second is at the level of scales or composites derived from the responses to all of the exercises in a subject area or subarea. Both types of trends will be analyzed. Exercises that are repeated in several waves of data collection give us the opportunity to see how the distribution of very specific knowledge or skill has changed over the years encompassed by the data. Scaled or composite scores derived from sets of exercises, which may or may not be repeated entirely in several years, will allow a more aggregate picture of the changes in the distribution of knowledge for each subject area across the relevant time periods. Trend analysis at the exercise level differs from that at the scale level in terms of both the appropriate questions to ask and the corresponding methods to apply.

Analysis at the exercise level. The question most appropriate for this level of analysis is: "How does the proportion of students who get the particular exercise correct vary over the years studied?" The main concern is to identify the overall trend across all students of a given age and also to identify significant student subpopulations exhibiting trends that differ from the overall picture. The overall trend is expressed by the "item × year" interaction while major subpopulations in which the trends differ will create a "subpopulation × item × year" interaction. These interactions may be analyzed most powerfully using the modern statistical theory of multi-way contingency tables (Bishop, Feinberg, & Holland, 1975).

By these procedures, one first forms a multi-way contingency table having at least these three dimensions: performance on the exercise (2 levels—right, wrong); year of data collection (4 levels); and subpopulation membership (n levels). Examples of subpopulations are sex, ethnicity, region of the country, urban-rural, and so forth.

Strictly speaking, the dimensions ought to include those that describe the sampling frames for each year. This permits one to use the unweighted data and simplifies the sampling properties of the relevant test statistics. In this framework, the overall trend in the proportion correct is associated with the item × year interaction as expressed in an appropriate log-linear model for the multi-way contingency table. Equivalently, logistic regression methods can be used to obtain the parameter estimates.

Serious deviations from the overall trend for the given exercise may be determined by testing for subpopulation × item × year in-

teractions using log-linear models for the multi-way table. This will result in identifying two classes of exercises that are repeated over time. The first type will be those exercises for which the time trend is fairly consistent across all major subpopulations. The second type will be those exercises for which there are significant differences in the time trends across subpopulations. The use of modern contingency table methods allows these two types of trends to be rigorously identified and distinguished from one another.

In addition to the time trends just described, further analytical power is afforded when the number of years intervening between assessment waves matches the age difference of the samples assessed. For example, the cohort of students assessed in 1979-80 at age 9 will be 13 years old in 1983-84. Similarly, 13-year olds in 1979-80 will be 17 years old in 1983-84. Exercises that are repeated in these two waves of data collection and which are administered to both 9- and 13-year olds or to 13- and 17-year olds give us a double-barreled look at time trends. We can investigate how a cohort, say 9-year olds in 1979-80, changed in their responses to a repeated exercise when the cohort became 13-years old in 1983-84, and we can compare these changes to that for other cohorts. The statistical tools for carrying out these analyses are similar to those described earlier.

Although such a linking of assessment intervals [illegible]ge differences in the sample occurs only sporadically in pa[illegible]sessment waves, appropriate matches do occur for writing (1969-70 and 1973-74), reading (1970-71 and 1974-75), social studies (1971-72 and 1975-76), science (1972-73 and 1976-77), art (1974-75 and 1978-79), and mathematics (1977-78 and 1981-82). If the schedule outlined in Table 1 is adhered to, appropriate cohort matches would be routine in the redesigned NAEP. In addition, using this proposed schedule, a cohort match occurs immediately in reading (1979-80 and 1983-84) and a full cohort cycle is achieved in mathematics (1977-78, 1981-82, and 1985-86).

Analysis at the scale level. Trend analysis at the scale level is concerned primarily with how the distribution of scale scores for a given subject area changes over time. The issue of trends that are the same across all subpopulations versus those that differ in different subpopulations also arises as it did for individual exercises. An analytic tool that is appropriate for this type of analysis is the use of linear models to investigate the main effects of year and the year × subpopulation interactions. The use of linear

models is geared for studies of changes in the means of the distributions. Studies of changes in other features of the distributions of scores are more appropriately done using plots of the data once the effects on the means have been isolated.

Reporting results. Once the significant results of the trend analyses are discovered, they will be summarized in simpler tables in which the data are weighted appropriately to give population estimates for the level of each variable (percentage or scale score) across years and, if necessary, across the relevant subpopulations.

"Causal" or Path Analysis

If NAEP is conceived mainly as a data collection function with a mission to develop and report population estimates of educational attainment for various groups over time and to codify the data on public use tapes for others to analyze, the enterprise is doomed to limited and sporadic impact. What is needed is a sustained program of analyses that seek reasons for the various levels of educational attainment and attempt to delineate their implications for policy alternatives. The availability of public use tapes will stimulate some of this activity by investigators throughout the country, but availability of data tapes alone will not sustain it. Every effort should be made to buttress widespread use of the data tapes because the ideological nature of education demands a multiperspective examination. One way to accomplish this is to maintain a continuing NAEP program of educational and policy analysis that would provide timely perspectives on emergent and recurrent issues and at the same time stimulate and facilitate other investigators to elaborate, modify, and challenge NAEP findings and interpretations.

This approach stresses analyses which focus on possible explanations of successful and unsuccessful performance. For example, that males outperform females in mathematics at a particular age may be a fact, but its policy and action implications would differ depending on whether there are also large sex differences in attitudes toward mathematics and in the number of mathematics courses taken. We do not contend that analyses of correlations based on nonexperimental survey data can answer questions of cause and effect, but such analyses can lead to rejection of some proposed explanations as inconsistent with the existing data and may suggest hypotheses for future survey mea-

sures and for formal experimentation by others in different settings. By this means, NAEP would not only report facts but relate them to context and to policy alternatives.

Background and program variables. To relate NAEP achievement data to issues of educational practice and policy requires additional information about the backgrounds of the students assessed and about their experiences in schools and programs. Some information of this type is already being collected by NAEP, but the coverage of the student and school questionnaires needs to be extended to allow us to address more fully the kinds of national concerns, human resource needs, and program effectiveness issues raised in Chapter I. Granted that questions to students and principals cannot be expanded indefinitely, but they can be expanded considerably beyond their current limits. Furthermore, much school and community information can be assembled by NAEP field personnel.

The variables to be tapped should be carefully chosen from a structured array of alternatives so that priority judgments are required and systematically justified (Messick & Barrows, 1972). These variables may differ from subject area to subject area, from age level to age level, and from assessment wave to assessment wave, but a core set of key common variables should be retained.

The kinds of student and background variables to be considered include demographic descriptors; nonNAEP measures of academic achievement; participation in special programs; measures of attitudes, interests, aspirations, and plans; of time spent studying, reading, viewing TV, in athletics and other activities, and (for older students) in employment; and, of a variety of family status and process characteristics. The kinds of school and program variables to be considered include school descriptors for racial, ethnic, and SES composition as well as desegregation history; size and type of school and community; availability of special programs; types of curricula, tracking arrangements, and extra-curricular activities; resource utilization; and, indicators of school climate and image.

In selecting specific variables, guidance would be sought from the educational literature but will be evaluated with great care. For example, measures of school facilities and curricula were only weakly related to verbal achievement in the Coleman (1966) equal educational opportunity survey. But the measures reflected neither the quality nor the utilization of the facilities and curricula, yet they still appeared to have more impact for some types

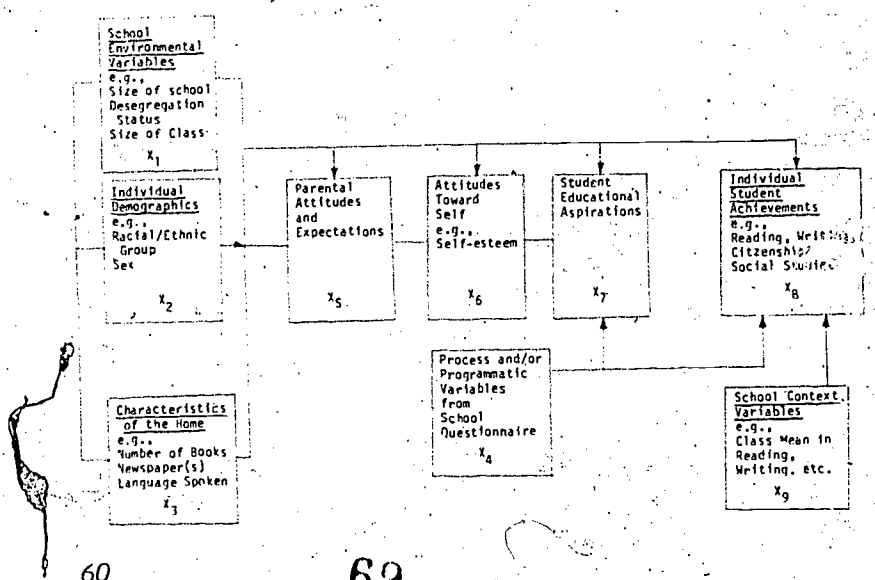
of students than for others in more refined analyses (M. S. Smith, 1972). Furthermore, the educational achievement criteria in the Coleman study were distinctly different from the subject-area exercises of NAEP.

Structural models and path analysis. Given that some amount of information will be available about student background, home and school environment, and program participation, structural equation or path models of educational attainment can be formulated and tested. Path analysis is a technique used to assess the direct or so-called "causal" contribution of one variable to another in nonexperimental data. The word "causal" is not meant to imply any deep philosophical connotation beyond a shorthand designation for an unobserved hypothesized process. The general problem is that of estimating the parameters of a set of linear structural equations representing the cause and effect relationships hypothesized in a particular theoretical conception.

Several recent path models incorporate unobserved latent constructs or factors which, while not directly measured, have operational implications for relationships among observed variables. In some models the observed variables are viewed as effects of the hypothesized constructs while in others they serve as causes, or as both causes and effects, of the latent constructs (Jöreskog &

Figure 7

Hypothetical Explanatory Model



Sörbom, 1979; Bentler, 1980). In effect, this approach combines path analysis with factor analysis.

As an example of structural modeling, a hypothesized explanatory model for student achievement is given in Figure 7. Individual student performances in reading, writing, and citizenship/social studies are hypothesized to be functions of a number of other variables including demographic, attitudinal, expectational, aspirational, and peer dimensions as well as characteristics of home and school environments and of school processes and programs. All of these components combine to form a network of specified interactions that affect educational performances. Indeed, educational performance in turn may affect some of its components such as aspirations and attitudes toward oneself. Simply to report differences in performance for different groups, while ignoring the available data for exploring this network, leads the recipient of the results to engage in uninformed speculations about their meaning and possible causes.

It is anticipated that explanatory models similar to Figure 7 will be formulated and tested both within and across population groups. For example, it is possible that the size of the relative effects and the processes through which they act—that is, indirect effects—may be different for different sex, race, ethnic, and age groups. Comparisons between models for 9-year olds and 17-year olds, as an instance, may suggest that school-related program variables have a steadily increasing impact while parental variables decrease in influence during this transition. Cross-ethnic and cross-sex group comparisons of similar models may be particularly informative with respect to how different programs and objectives affect such subgroups.

Past analyses of educational performance have faltered on a variety of technical problems. The traditional approach, as exemplified by the Coleman equal educational opportunity survey, used a single equation model of educational attainment and employed regression analysis to estimate the degree to which different components affected achievement. Such an approach has no way to disentangle the correlations among predictor variables. Since the order in which variables are entered into a regression equation markedly affects the estimate of their importance and since Coleman entered school variables last (thereby minimizing the estimate of their effects), it is little wonder that he concluded that schooling had little impact. Later investigators of the same data using different analytic methods showed a substantial effect

of school variables (Mayeske, Wisler, Beaton, Weinfeld, Cohen, Okada, Proshek, & Tabler, 1972).

Recent developments in path analysis and in the analysis of structural equations (Jöreskog & Sörbom, 1979; Bentler, 1980) make it possible to specify much more realistic explanatory models of educational performance and to avoid some of the technical problems of regression analysis. Basically, the network of relationships in the explanatory model is represented by a set of equations, and the data are used to estimate the unknown coefficients of the equations and the degree of confidence that can be placed in the estimates. A very flexible computer program, LISREL V (Jöreskog & Sörbom, 1981), is available for the computations.

Several advantages of using structural equations should be noted. Parameters for the entire model are estimated simultaneously, thus avoiding the bias involved in estimating the equations separately by regression analysis. Reciprocal relationships may be introduced, such as the effect of performance on attitudes as well as the effect of attitudes on performance. The explanatory variables in the model need not be considered to be measured without error, as in regression analysis. Furthermore, the errors in the variables may be assumed to be correlated. When two or more variables are combined into a composite, a reliability is computed, reported, and used in the estimation procedure.

Special Studies

Inevitably, a number of special concerns arise over the years that NAEP cannot readily address within its regular financial resources but that would be beneficially addressed within the NAEP environment. This is because the special studies, if done in the NAEP context, might be tailored to benefit NAEP functions or broaden its purposes, while at the same time the study in question capitalizes on existing facilities or ongoing activities. For these reasons, NAEP should be committed to a continuing effort to develop funding for such additional studies from private foundations or appropriate government agencies. The following kinds of studies should be high on the agenda.

Assessment of Functionally-Handicapped Students

In a recent report of the National Academy of Sciences (Heller, Holtzman, & Messick, 1982), the educational progress of educable mentally retarded and other functionally-handicapped students was singled out as the touchstone for equity in special education. Since such students are currently excluded from NAEP, it seems fitting that NAEP attempt to mount a special assessment of their educational competencies and ultimately of their educational progress. Indeed, the effort would be facilitated if such students were not only identified for exclusion in the NAEP sampling process, but were described in more detail in regard to their background and program experiences, as proposed in this NAEP redesign.

Assessment of the competencies of handicapped students faces a number of major roadblocks because of fundamental problems in exercise development, administration, and interpretation that are encountered (Bennett, in press). The Education for All Handicapped Children Act (PL 94-142) requires that educational goals for handicapped students be individually prescribed. From the standpoint of assessment, this requirement results in the creation of an unmanageably large array of goals from which common objectives for exercise development may not be easily extracted (Maher & Bennett, in press). The diverse needs of handicapped students also demand departures from traditional exercise formats; exercises ordinarily printed in standard form must typically be created in braille, cassette, and large-type versions. Administration is made difficult because many disabled students require untimed individual administrations. Special probes and monitoring may also be required to assure that the instructions are understood.

Such departures from standardized conditions, as well as the required variations in exercise format, in turn create dilemmas for data interpretation. Aggregation of data is at best problematic and at worst pointless unless individual assessments can be placed on a common scale—or unless some kind of defensible basis for comparability can be realized. However, if comparable assessments can be achieved for students with the same type of handicap, then their educational progress could be monitored even though it would not be strictly comparable to the progress of other handicapped or nonhandicapped groups.

These difficulties were recited not to justify exclusion of this important segment of the school population from assessment, but to underscore the nature of the challenge to measurement specialists and to make it clear why this effort should be a series of special studies rather than an integral part of NAEP. To begin with, the problems of assessing the mentally and physically handicapped require concentrated attention and a full-scale attack that should capitalize upon the NAEP field presence in schools but should not disrupt that presence or the regular NAEP activities. Ultimately, if these assessment problems can be solved, the educational progress of functionally-handicapped students might become an integral part of the national assessment.

Assessment of Limited-English Speaking Students

Since the other major group of students excluded from NAEP—the nonEnglish proficient—typically come from ethnic minority groups, their continued exclusion may seriously bias interpretations of the educational progress of those ethnic groups. Furthermore, as with special education for the handicapped, the touchstone for equity in bilingual education is the educational progress of the students. For these reasons, NAEP should mount a special study attacking the measurement and logistical problems in assessing non-English proficient groups.

These problems are no less formidable than those of assessing the handicapped. First, exercises must be developed in a number of different languages—although this might be addressed in waves of one language at a time, beginning with Spanish because of the size of the Hispanic minority in the country. Aside from the substantial resources required to accomplish this, differences among languages make it difficult to develop non-English exercises that are precisely comparable to English-language versions. Second, non-English proficient students often vary in their knowledge of the written form of their language. Even though they may speak that language better than they speak English, they may not read that language well enough to be examined in it via printed exercises. This underscores the point that one of the goals of assessing limited-English speaking students should be assessment of their proficiency in both English and their native language. Finally, inclusion of students from backgrounds providing

little preparation for formal examinations necessitates using specially-trained examiners to assist students in understanding the requirements of the examination situation.

Again, this litany of troublesome problems is not meant to justify continued exclusion of non-English proficient students from NAEP, but to highlight the need for a special frontal attack on an important national issue in educational assessment.

Innovative Exercise Development

Although attention to innovative exercise development should be a routine part of NAEP's day-to-day activities, the focus in that context tends to be on the development of new ways—that are more valid or efficient or interesting—to measure dimensions already being measured in old ways. In contrast, this proposed special study focusses as well on the development of new ways of using old methods to assess new dimensions and, most importantly, on new ways of assessing new dimensions that have previously been difficult to capture. It is proposed as a special study because a critical mass of attention and effort is needed at the outset, although the innovations developed and the innovative mode of development should ultimately be incorporated as standard NAEP approaches.

As an example of using old methods in new ways to measure new dimensions, consider the possibility of using integrated sets of multiple-choice items to assess complex problem-solving or decision-making processes in a subject area. Since each step in complex problem solving entails a decision point or a set of decision points, multiple-choice items could be constructed to assess the choices made—for example, the kinds of information sought, the strategies utilized, the hypotheses generated, the analyses undertaken, the alternatives weighed, the solutions selected, and so forth, perhaps each with an associated item that requires selection of the reason for each move. The multiple-choice formats would be broadly conceived to include matching and keylist procedures, for example, as well as more standard versions. Such integrated sets of exercises could also be branched depending upon the choices made at each point, with or without provision for recycling.

As another instance, if multiple-choice exercises were constructed so that selection of incorrect distractors were indicative

of common errors made during learning, then patterns of distractor choice might be diagnostic of instructional problem areas. With such exercises, reports of average percent correct could be accompanied by summaries of the types and frequencies of errors made, thereby enriching the utility of the results for instructional purposes at the classroom level.

Both of these examples illustrate a means of overcoming one of the major criticisms of multiple-choice exercises—namely, their rigidity of application and orientation to outcome rather than process. At the same time the new uses retain the major advantages of multiple-choice methods—namely, the economy, efficiency, and ease of administration and scoring that historically have tipped the scale in favor of their use over other types of exercises.

An example of new ways of assessing new dimensions that have been elusive in the past is the use of problem simulations, which might be presented by printed material or by film or videotape techniques. Students might be asked to generate as many alternative hypotheses as they can for a given problem, for example, or as many alternative reasons as they can for a given outcome. Such productive responses could then be judgmentally scored for fluency, flexibility, and originality or other aspects of divergent thinking (e.g., Frederiksen & Evans, 1974; Ward, 1982; Ward, Frederiksen, & Carlson, 1980). The simulations could also be constructed to assess sensitivity to problems or problem-finding skills in various subject areas.

With videotape technology, simulated interpersonal scenes could be presented and periodically interrupted with questions or tasks to assess sensitivity to interpersonal cues, appreciation or tolerance of individual and group differences, and a variety of other social skills (e.g., Stricker, 1982). In addition, videotape presentation could facilitate assessment of understanding and appreciation of the performing arts. Finally, computer technology offers another powerful vehicle for innovative exercise development which will be briefly discussed below.

Computer-Assisted Assessment

Available computer technology can improve the efficiency of a number of NAEP activities almost immediately—such as the use of computer networks for remote conferencing, which would facilitate committee work on such activities as objective setting and

exercise review while reducing the number of face-to-face meetings required. Another instance is remote access to NAEP data bases for special analyses or inquiries by the various NAEP committees or by NIE. If such capabilities have not yet been introduced, they should be explored in the near future. However, direct contributions of computer technology to the main NAEP activity of assessment require special study. Such a special study or set of studies should not only address the feasibility and appropriate timing of introducing computer-assisted assessment into NAEP, but should attempt to develop the technical means for optimizing computer use in exercise development and administration.

One set of issues involves the use of the computer for exercise administration—such as to insure proper spiralling of exercises within and across subject areas during individual administrations or to obtain efficient assessments of individual skill levels via tailored-testing procedures (Lord, 1977, 1980b), or some combination of both. Another set of issues involves the development of measurement procedures and innovative exercises that capitalize on the algorithmic and heuristic capabilities of the computer to improve the assessment of existing and new skill dimensions. For example, with computer administration, latency and speed measures could be routinely obtained which might prove of value in the assessment of mastery in reading, computation, and other performance skills; such measures applied to knowledge retrieval exercises should also buttress the assessment of subject mastery.

In regard to new skill dimensions not well covered previously, the computer makes possible the assessment of information processing skills that are difficult to assess by other means—such as skills involved in information search and organization, hypothesis generation and testing, restructuring of information, and other components of complex problem-solving and decision-making tasks or other types of sequential thinking. This is possible because the computer can record the paths, speed, and outcomes of such activities as they occur on subtasks within the sequence —in contrast to the limited and schematized attempts discussed earlier to mimic this process with integrated sets of multiple-choice exercises.

Special studies were highlighted in this NAEP redesign because some ongoing capability to probe and explore important assessment and development opportunities is needed as a basis for NAEP's continuous improvement and renewal.

III.
Enhancing NAEP'S Flexibility To Meet Varied Assessment Needs

The proposed NAEP redesign affords vast flexibility in data analysis and in relating data to a variety of policy issues. But sophisticated analysis is not enough—in addition, NAEP needs sophisticated ways of communicating the results and of targeting the presentations to the needs of various audiences. Furthermore, NAEP's capacity to meet a variety of assessment needs would be markedly enhanced by linking NAEP data to other national, state, and local data sources and by extending refined NAEP services to a broader clientele. Finally, since the objective-setting process is just one step removed from the standard-setting process and since NAEP results bear directly on attained performance levels, NAEP should actively confront the issue of educational standards—not to set them, but to clarify them and to help the various interested publics to set their own standards. Each of these points is briefly discussed in turn in the ensuing sections.

Flexibility in Analysis and Reporting

We have seen how the availability of covariances among exercises as well as the availability of scales having common meaning across population subgroups, age levels, and time periods serves to improve the meaningfulness and interpretability of assessment results and trends. These are among the most important of the benefits deriving from BIB spiralling and IRT scaling, but they are by no means the only important benefits. We next review how IRT scaling provides great flexibility in relating achievement data to policy questions. We then review methods for flexibly presenting achievement data so that its meaning and import are readily revealed in a particular policy context or to constituencies with particular concerns.

Responding to Multiple Policy Issues

An important by-product of the IRT scaling of NAEP exercises is that estimates are available of each respondent's skill levels for those areas in which he or she was assessed. This means that the various achievement dimensions scaled by IRT may be correlated with any of the variables of background, attitude, school, and program that are tied to those individuals via the student and school questionnaires, school records, or other means.

Furthermore, these variables could also be used to generate group comparisons—such as students in college preparatory versus vocational programs, students in private versus public schools, or students exposed to preschool programs versus those who were not. Although the resulting sample sizes in many of these group comparisons will not be large or nationally representative, they may be sufficient to provide timely provisional answers pending more intensive investigation. Given the availability of other background variables characterizing the groups in question, these group comparisons may also be conducted controlling for a variety of home, school, and demographic factors by means of analysis of covariance techniques. Although student skill estimates are not reliable enough for reporting at the individual level, they are sufficiently reliable for comparisons at the group level as well as for correlational work—where in any event unreliability can be taken into account.

The only limitation on the nature and number of educational and policy questions that can be addressed in this fashion is whether or not relevant background and program variables were included in the student and school questionnaires or are derivable from other sources. The capacity to respond to new policy questions with existing data thus depends on our luck or our wit in having included variables pertinent to the questions.

Communicating Results to Multiple Audiences

The most effective way to communicate complex statistical results is with graphical formats (Wainer & Thissen, 1981). Paradoxically, one rather compelling bit of evidence supporting this is the often poor quality of published graphics. The continued existence of poor graphics is partially due to the amazing capacity of a human audience to be able to understand graphs accurately and

quickly even though they contain serious logical or technical faults. This helps to explain why so many of the empirical investigations into the efficacy of various graphical formats have shown variable results and small differences in efficacy among the alternative forms of graphs (MacDonald-Ross, 1978; Wainer, Groves, & Lono, 1978, 1979). This tends to be true for large effects in simple data structures, however, where any reasonable display will work. When the effects are subtle or the data are complex, the displays must be done wisely.

Graphics both clarify and reveal relationships. To illustrate how a good display can provide still more information after it is redesigned, consider the data in Table 4, which originally appeared as Table 10 in the 1981 NAEP Report Number 11-R-01. The table presents all the information required to see certain effects—most notably the increase in performance of the lowest achievement class of nine-year olds. We note that the data given in Table 4 are distributional, providing achievement summaries for the various ability levels in each of three birth cohorts. To show these distributions more clearly still, we can utilize a variant of a box-and-whisker plot (Tukey, 1977). We will use a dot to represent performance in the extreme ability groups, and horizontal lines to represent performance in the other two groups as well as a heavier line to represent the national mean. These horizontal lines will then be connected to form boxes which enclose approximately the middle 50% of the students. Such a plot is shown in Figure 8.

The display in Figure 8 forces us to see what we had to look closely for in Table 4—specifically, we note that among the nine-year olds the lowest achievement group is further from the rest than appears to be the case in the other age groups. Also, an increase in performance in 1981 (the 1971 birth cohort) is evident, especially in the lowest achievement group. An interesting facet of these data revealed in this plot is that the 1962-63 birth cohort seems to perform more poorly than the other birth cohorts. This is seen in the 9-year old data (where those 9-year olds born later do better) and again in the 17-year old data (where those 17-year olds born earlier do better). Thus, we begin to see some longitudinal characteristics from these cross-sectional data. Our ability to observe these interesting effects is partially due to the display methodology. Note that notched box plots (McGill, Tukey, & Larson, 1978) could also be used to provide visual information on the statistical significance of observed visual differences.

Table 4

National Results by Achievement Classes: Mean Percentages and Changes in Correct Responses for Ages 9, 13 and In-School 17 on Inferential Comprehension Exercises in Three Reading Assessments†

	Age 9: 27 Exercises					
	1971	Change 1971-75	1975	Change 1975-80	1980	Change 1971-80
Nation	60.5%	0.9	61.4%	2.5*	63.9%	3.5*
Achievement class 1	35.5	3.3*	38.8	4.7*	43.4	7.9*
Achievement class 2	57.8	1.5	59.3	2.2*	61.6	3.7*
Achievement class 3	68.5	-0.1	68.4	1.4	69.8	1.2
Achievement class 4	80.0	-0.8	79.2	1.8	81.0	1.0

	Age 13: 24 Exercises					
	1970	Change 1970-74	1974	Change 1974-79	1979	Change 1970-79
Nation	56.1%	-0.8	55.3%	0.2	55.5%	-0.6
Achievement class 1	35.0	1.2	36.2	0.5	36.7	1.7
Achievement class 2	50.8	0.1	50.9	0.9	51.8	1.0
Achievement class 3	61.8	-1.3	60.6	-0.4	60.2	-1.7
Achievement class 4	76.6	-3.1*	73.5	-0.4	73.1	-3.4*

	Age 17: 25 Exercises					
	1971	Change 1971-75	1975	Change 1975-80	1980	Change 1971-80
Nation	64.2%	-0.9	63.3%	-1.2	62.1%	-2.1*
Achievement class 1	39.1	2.5*	41.6	-1.4	40.1	1.0
Achievement class 2	58.7	-0.1	58.6	-2.0	56.7	-2.0
Achievement class 3	72.3	-2.6*	69.7	-1.2	68.4	-3.9*
Achievement class 4	86.8	-3.4*	83.5	-0.3	83.2	-3.7*

†Figures may not total due to rounding.

*Indicates significant change in performance between assessments.

Note: Achievement class 1 = lowest one-fourth
Achievement class 2 = middle lowest one-fourth
Achievement class 3 = middle highest one-fourth
Achievement class 4 = highest one-fourth

Cohort effects are seen only by contrast with these data because the dependent variable (percent correct) cannot be compared across age levels—that is, 46 percent correct in the assessment of 9-year olds does not compare to 46 percent correct in the assessment of 13-year olds. Yet, if the exercises were linked or equated in some way, we would be able to make these kinds of comparisons. Using the IRT scaling methodology espoused in this proposed NAEP redesign would yield an underlying skill scale on which all groups could be directly compared. A plot of how such hypothetical data might appear is given in Figure 9.

Figure 8

National Results by Achievement Classes: Mean Percentages and Changes in Correct Responses for Ages 9, 13, and In-School 17 on Inferential Comprehension Exercises in Three Reading Assessments

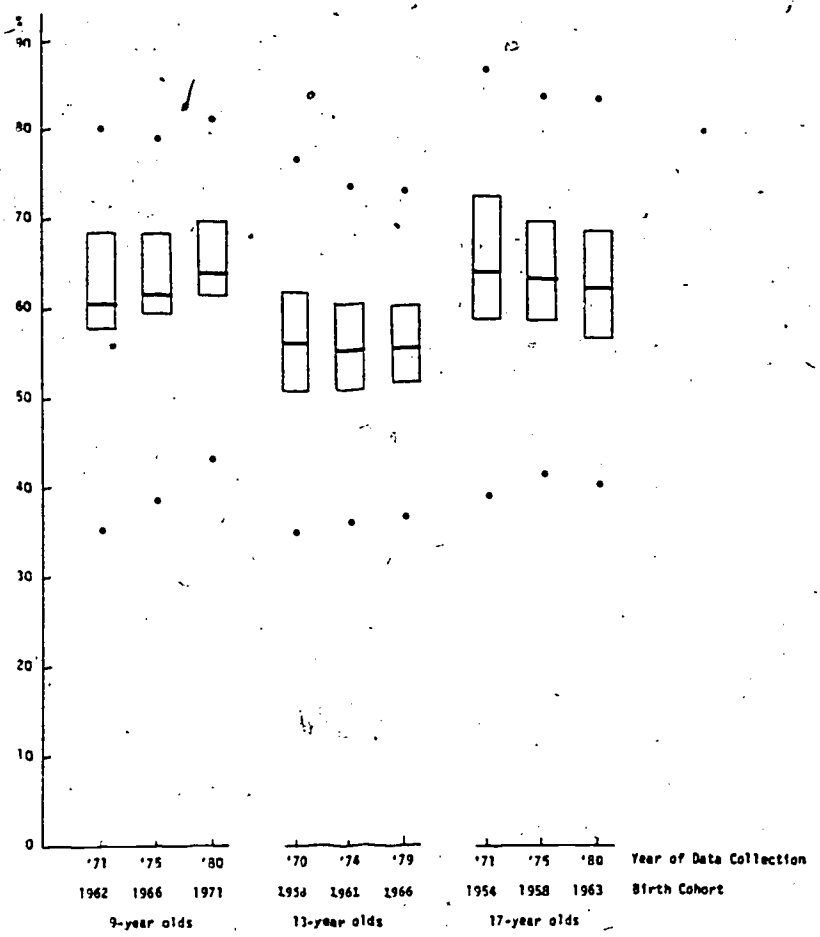
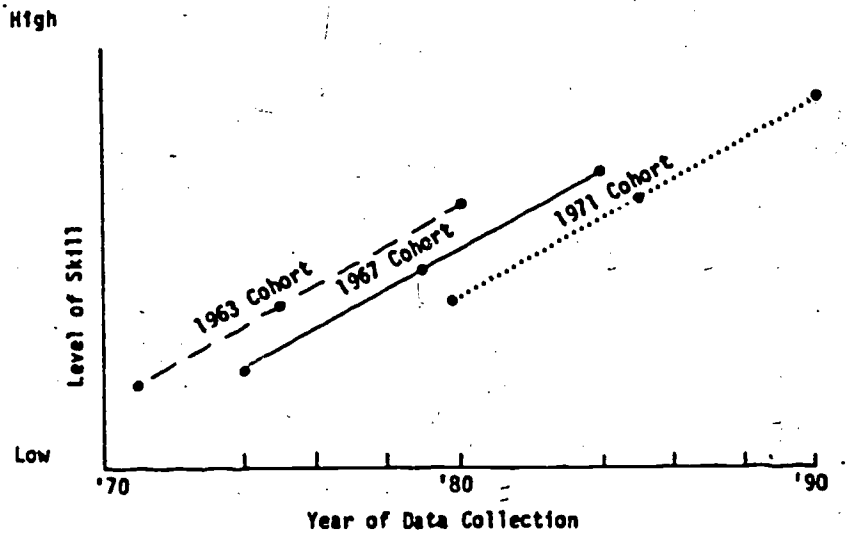


Figure 9

Hypothetical Example Showing Longitudinal Trends Within Cohort



The shift in emphasis from Figure 8 to this plot is the connecting of birth cohorts over time. The slope of these connecting lines provides a measure of the rate of educational growth. The location of the points provides a measure of the change seen across cohorts. In Figure 9 we see increases in skill from the 1963 cohort to the 1967 cohort to the 1971 cohort. If such a finding did occur, we might then look to exogenous variables to provide explanatory clues for the upward migration—such as better instruction, increased emphasis on basics, or newer teaching techniques. If desired, one could use box plots rather than points to provide a fuller picture of change in the entire distributions of skill.

It should be clear from this example that explaining complex data structures in prose or in tables provides neither the ease of comprehension, nor the richness of interpretation, that is available in even these straightforward plots. More complex data require still more imaginative plotting techniques—for example, how would one show the same results as in Figure 9 broken down by geographic region?

Proposed graphical reporting system. With these illustrations in mind, we can better discuss the proposed approach to the

reporting of results. This approach is principally graphic in orientation, flexible in design, and takes advantage of the latest computer technology for the plotting and dissemination of findings.

Since NAEP results are of interest to a diversity of audiences, trying to provide a single report or reporting mode that would satisfy all interests is doomed from the start. Trying to anticipate the various audiences and providing parallel documents at each of these levels has a greater possibility of success, but it is a very difficult and cumbersome chore. What seems a more fruitful approach is to provide information at a variety of levels for what are clearly the major audiences, yet simultaneously have the capability for quick and easy generation of graphical and tabular answers to questions asked on an ad hoc basis. Thus, one might have a general answer pre-prepared for salient questions (e.g., What is the mean reading ability of the 1961 birth cohort from age 9 until age 17?) and allow specialized answers to be generated on demand (e.g., the same question, but just for rural schools).

What is needed is an interactive dynamic system of graphical data analysis to provide the capacity both to make quick responses and to ask questions suggested by the answers to previous questions. This system should provide both static and kinematic display capabilities. For communication in the traditional print media *static displays* continue to provide an accurate and efficient method. Wisely chosen graphs can often deliver quite complex messages. We expect that this will continue to be the principal mode of information dissemination using computergraphic hardware and software linked to the appropriately structured NAEP data base. Recent developments in computer technology and software design also facilitate the routine use of *kinematic displays*, which make possible compelling and informative data presentations via film or TV media.

Kinematic displays have a number of overlapping uses. First, an interactive kinematic display provides an easy way for an investigator to explore both the gross and fine structure of complex data, by panning around the data structure noting regularities and then zooming in on irregularities and outliers. Using such techniques one can spot an unusual data configuration, zoom in on it and immediately bring to bear exogenous program or background variables to try to understand plausible causes for the atypical behavior. Second, complex multivariate data structures are often best seen in a kinematic display. The precise sort of display depends on the data. For example, with three-dimensional data, one can

produce an evocative three- onal image by rotating the three-dimensional scatter p l time. Even though the display is on a flat screen, the i rceived is three dimensional.

Another kind of kinematic display, which is quite useful for viewing and comparing a series of two-dimensional figures, is the alternagraphic plot. This method alternates two or more plots which are to be compared quickly enough so that the eye superimposes one on the other, yet slowly enough so that the separate displays can be seen as well (about 500 milliseconds each).

As a quick illustration of how NAEP might use some of the simpler aspects of this kinematic display technology, consider some variations on Figure 9. Suppose we were interested in comparing the data shown in that figure with the same data for a specific subpopulation such as an ethnic, sex, or regional breakdown. We could use an alternagraphic display, alternating back and forth between the data in Figure 9 and the data for the subpopulation. A short viewing time would provide a clear picture. This method could be expanded to more than two plots.

While kinematic displays provide a powerful data-analytic tool for investigators, the main intention here is for the communication of results to a broad audience. The vast majority of the U.S. population get most of their information about the outside world through the video media. Thus, in order to communicate facts and understanding about complex data structures to the public, it would be a matter of small difficulty to prepare video tapes using kinematic display technology. The possibilities opened up by such a capacity are both broad and exciting; the time is certainly ripe for their exploration and use.

Extending NAEP's Impact

The impact of NAEP results could be both extended and enriched by linking NAEP data to that in other data bases and by linking the national assessment program to other assessment programs.

Linking to Other Data Bases

The power and value of past and future NAEP data would be tremendously enhanced if the responses of NAEP samples could be di-

rectly compared to, or interpreted in the light of, the responses of different samples to the same or demonstrably equivalent exercise materials. For example, the use of NAEP exercises in other national surveys could provide trend data not otherwise available given the spacing between assessments. Furthermore, the use of NAEP exercises in samples with a different design—perhaps a national sample in which multiple minority groups have been systematically oversampled—would permit more intensive investigation of differential performance correlates than is possible with NAEP data alone. In addition, there is also the possibility of linking NAEP findings to data bases in which more comprehensive descriptors of the respondents are available. These data might include extensive student variables (cognitive and noncognitive), background factors (ethnic, parental), or situational characteristics (school, community, labor market).

These linkages could come about in three major ways: (1) by use of NAEP exercises in other surveys where the data collection procedures were sufficiently similar to permit comparisons; (2) by equating NAEP exercises to similar measures in other assessments and surveys; and, (3) by embedding NAEP exercises in the instrumentation for future assessments and surveys. Each of these possibilities is briefly discussed in the ensuing paragraphs.

NAEP exercises in other surveys. Since NAEP exercises were developed with great care and the associated response data provide a national perspective, it would be beneficial to use released NAEP exercises in other surveys. Indeed, this was done in the 1980 data collections of High School and Beyond (HSB)—the name given to the new high-school cohorts surveyed in the spring of 1980 in the national longitudinal studies sponsored by the National Center for Education Statistics.

In 1978, the test battery for High School and Beyond was designed by Educational Testing Service. Wishing to include in the battery a set of exercises measuring science knowledge, ETS recommended that NAEP exercises be used in order to fulfill several objectives, one of which was the establishment of a link between NAEP and HSB. The NAEP science exercises were included in the 1980 sophomore battery. Then, in 1982, the original sophomores were given exactly the same science exercises again, at which time most of the students were seniors.

Because of differences in the mode of administration, NAEP and HSB data on these same science exercises differ in a number of ways. In NAEP the exercises were group administered with tape-re-

corded instructions and generous time limits, whereas in HSB the instructions were read and explained by a survey administrator and there was a 10-minute time limit for 20 science questions. Moreover, the NAEP exercises had six options including "I don't know", whereas in HSB the "I don't know" option was omitted. Also, it should be kept in mind that the NAEP cohorts were selected by age, whereas the HSB respondents were grouped by high school grade level.

Thus, even when the respondents are comparable as far as educational development is concerned, there are some possibly serious constraints on what can be concluded from comparisons between the performance of NAEP and HSB samples. But there may also be some useful comparative findings as follows:

(1) Since the HSB respondents who first took the NAEP science exercises as sophomores later took the same exercises as seniors, the HSB results provide some useful data as to which exercises are the most sensitive measures of growth in science knowledge from the sophomore to the senior year. Also, since the HSB data are certain to be used in studies of school effects, there should be information on the correlation between the science exercises and school variables.

(2) The HSB data file has a much broader range of information on the characteristics of individual students and on the schools they attended than does the NAEP file. The HSB file thus provides a more comprehensive picture of the characteristics of students who were successful on the NAEP science exercises in comparison with those who were not successful.

(3) Since scores on the Scholastic Aptitude Test and on the Armed Services Vocational Aptitude Battery are being retrieved for HSB students, it would be possible to link performance on the NAEP science exercises to performance on the SAT and ASVAB.

(4) As part of an evaluation of the HSB battery, the sophomore data were factor analyzed, with the results reported in Table 5 (Heyns & Hilton, 1982). These results suggest that the NAEP science exercises, as administered under HSB conditions, reflect a set of fairly broad cognitive abilities—as witness the science loading of .61 on a verbal factor and .21 on a math factor, with some variance left over reflective of science information.

Other linkages to existing data files are possible and may provide valuable insights. Approximately 25 states have used NAEP exercises in various numbers and in various ways (usually in large group administrations). As with the HSB data, the state data could

Table 5

Two Factor Solution for "High School and Beyond" Sophomores and Reliabilities (N = 26,110)

	Factor Loadings		Percentage of Variance Accounted For	KR 20
	Verbal	Mathematics		
Vocabulary	.83	—	68	.81
Reading	.86	—	74	.78
Mathematics I	—	.94	88	.85
Mathematics II	—	.72	52	.54
Science	.61	.21	64	.75
Writing	.61	.18	60	.80
Civics	.69	-.01	45	.53

	Correlation Between Factors	
	V	Q
Verbal	1.00	0.841
Quantitative	0.841	1.00

be particularly valuable where relatively large numbers of special populations were tested, a possible example being Native Americans in certain western states. As a final example of the inclusion of NAEP exercises in other surveys, we mention the possibility of foreign administrations, which would provide the opportunity for an international perspective on educational achievement. These could be programmatic cross-national surveys, such as the international studies of comparative educational achievement conducted by the International Education Assessment, or cooperative arrangements for the exchange of exercises with the national surveys of other countries.

Equating NAEP exercises to other existing measures. Where the interest is in linking NAEP exercises to similar but not identical exercises already used in other surveys, it may be possible to equate the two sets of exercises by means of specially designed equating experiments. As an instance, Beaton, Hilton, and Schrader (1977) equated similar exercises from two different data sets as part of a study of the SAT score decline.

Some examples of relevant data bases that might be linked to NAEP via equating are Project Talent, the Coleman Equal Educa-

tional Opportunity Survey, the ETS Study of Academic Prediction and Growth, the NCES National Longitudinal Studies, the ETS-Head Start Longitudinal Study of Disadvantaged Children, and the Department of Labor National Longitudinal Survey.

Embedding NAEP exercises in future surveys. What is of considerably more promise is the possibility of embedding NAEP exercises in future surveys and in educational achievement tests developed by commercial publishers. On this latter score, a NAEP service offering commercial publishers an opportunity to obtain nationally-normed exercises would both upgrade the quality of educational testing generally and provide much needed revenue to NAEP for underwriting other activities. As a consequence, since commercially published educational tests are widely used in state and local assessments, the inclusion of NAEP-normed exercises embedded within them would both link these assessments to NAEP for purposes of research and provide a current national perspective for interpreting the state or local findings.

Extending NAEP Assessment Services

The ultimate value of NAEP must be viewed in terms of its contributions to a variety of users attempting to address important educational issues. Congressional appropriations as well as administration support for NAEP *assume*, and have a right to depend upon, optimization of these annual expenditures. Perfection of an instrument designed to yield specific reports to limited audiences can hardly be justified in today's political and economic environments. Thus, it seems reasonable for NAEP to pool resources with other interested parties for mutually advantageous purposes.

For example, asking states to share the costs of exercise development will both permit NAEP to do a better job and assure the state that high quality exercises will be available on their schedule at a fraction of what it would cost to develop them independently. Charging a state or a large city a $5,000 consulting fee for technical assistance might help it save $50,000 in expensive failure, while permitting NAEP to maintain a valuable service. Setting a reasonable fee to participate in a Large Scale Assessment Conference challenges NAEP to prepare a worthwhile agenda and at the same time discourages casual attendance.

One of the most important user groups is represented by the over 40 states that currently have some form of assessment or

testing program. It is not envisioned here that NAEP would provide services that are directly competitive with commercial or other non-profit organizations. On the other hand, it is possible to conceive an array of arrangements developed to accommodate states or large systems with or without a third partner.

For example, three assessment "packages" could be developed and made available to states to form part or all of their state assessment program. The main features of these "packages" would be that they

- provide a relationship to objectives and standards,
- permit comparison of state performance with NAEP national results,
- represent real cost savings to state assessment programs by providing already developed items of high quality and of known performance,
- include local options for specialized objectives, and
- replace expensive state-wide programs with an economical, high quality program, tailored to the state's needs and with results that permit comparison to national data.

These packages would be designed so as to be incremental—for example, as a first step, a state might contract with NAEP to provide exercises on a regular schedule for certain specified curriculum subjects. This would obviate the necessity for the state to develop its own test development capability. A second step might be for a state to contract for the complete test development process. A final step might be for the state to ask NAEP to run its complete state program simultaneously with the national data collection effort and provide the state with results and analyses.

The size of the state population assessed and the complexity of the program would impact costs, but in every case economies of scale should operate in favor of this being a less expensive alternative than a state managing a completely parallel effort. In addition, it may be found that samples for the national assessment and the state assessment can be drawn in such a way that they complement each other, to the mutual benefit of both assessments. In all of these versions, comparisons with national data would be possible.

In the past, arrangements with NAEP have been difficult for states because of postponements caused by NAEP budget changes and such. What is suggested here are contractual arrangements

with states which are quite independent of NAEP assessment budgets. As more states participate in this type of arrangement, the total program would be strengthened. It would obviously have to realize economies and greater quality for the states as well as income and facilitation of data collection for the national effort. The goal is a financially viable national assessment program, which would mean more innovative exercise development activity, more sophisticated data analyses, and more useful reports to school districts, states, government agencies, and the public.

Progress Toward Standards As Standards for Progress

The overall activities of NAEP skirt all sides of the issue of educational standards without addressing the heart of the matter. Most of the elements intrinsic to the setting and monitoring of educational standards are already an integral part of NAEP. These include the setting of learning objectives, the development of measurement procedures specifically geared toward those objectives, and the reporting of student performance levels in pursuit of those objectives. What is missing is a pluralistic process for taking the next step—for helping the various interested segments of society make the value judgments needed to set their own standards and to monitor and revise them over time. Descriptions of objectives that are commonly agreed upon and of performance levels that are currently being attained in different societal subgroups go a long way toward informing the societal standard-setting process.

Objectives and Standards

An important feature of NAEP's procedures is that the learning objectives guiding the assessment are determined by consensus as to their relevance and importance. This step is more than half the battle in standard setting because these objectives, in essence, are operational statements of what is worth teaching and important to learn. In effect, these objectives specify the areas in which it is worthwhile having [illegible]

A cautionary note is required here, however, because the concept of standards in a pluralistic society requires some provision for local variation and self-determination. In contrast, the principle of consensus might yield a common denominator that omits important educational goals not shared by everyone. Although a "national" assessment might reasonably be limited to common goals, it would not truly be national for a pluralistic nation.

What is needed is a method for augmenting the present system in order to obtain judgmental data descriptive of varying patterns of educational priorities set by different societal subgroups across the full range of objectives. Thus, by placing objective setting in the context of pluralistic standards, some of the pressure toward consensus would be relaxed. As a consequence, the total set of objectives would include not only those for which substantial consensus was achieved, but also those important objectives primarily embraced by substantial subgroups. Although different reporting profiles for different groups could be developed, it should prove more useful for each group to appraise performance levels on its own priority objectives in the context of the diverse objectives of other groups as well as in the context of the common objectives cutting across groups. Diversity of objectives is also the best protection against the elevation of consensual objectives to the level of implicit national standards.

For these reasons, it appears that objective setting should be addressed in the arena of pluralistic standards. Accordingly, we propose that the Exercise Development Committee of the Assessment Policy Committee be broadened to one on Objectives and Standards, with the charge not only to relate inwardly to the NAEP exercise-development process but to relate outwardly to the societal standard-setting process.

Performance Levels and Standards

Another critical element in the standard-setting process is information for each objective on the current performance levels and trends in various societal subgroups. Inverting the customary prescription that one must first determine the objectives of instruction before developing measures of learning outcomes, Henry Dyer (1967) once suggested that it might not be possible to decide what the objectives ought to be until one knows what the current outcomes are. The point is even more appropriate when applied

to standards. It might not be possible to decide what the standards ought to be until one knows what current performance levels are.

Detailed information on this point is available through NAEP, but it would be even more valuable if it were provided in the conjoint criterion-referenced and norm-referenced form made possible by IRT scaling, as proposed in this NAEP redesign. As summarized in Figure 5, IRT scaling permits one to estimate the proportion of correct answers to each exercise expected for individuals in each subgroup at any point on the skill scale. One can also estimate the proportion of individuals in any group who have less than some specified probability of success on any given exercise. This type of detailed information, as aggregated in various ways, provides the kind of group performance distributions needed to inform the standard-setting process.

Better still, the capacity to relate this group performance to scales anchored by benchmark exercises provides concrete exemplars for characterizing different performance levels. Eventually, the development of behavioral anchors for these dimensions, such as those exemplified by the Foreign Service Institute scale of foreign language attainment, would enrich this characterization with verbal summaries of related real-world capabilities associated with each scale level. What is still needed to move on to educational standards are the value judgments as to which performance levels are deemed unsatisfactory, adequate, or excellent by different societal groups.

Values and Standards

Our intent in broaching the issue of educational standards is not to involve NAEP directly in the standard-setting process, nor to settle for its indirect involvement as a mere data resource on consensual objectives and performance levels. As we have seen, NAEP is already directly involved in one critical aspect of the standard problem—namely, the choice via objective-setting of those areas that are worth teaching and learning and hence are worthy of standards. Since in making such choices NAEP needs to be sensitive to the pluralistic values of various societal groups, it seems sensible that NAEP should be more actively involved with societal groups on the issue of standards.

Again, the intent is not for NAEP to engage in the standard-

setting process, but to engage the public with NAEP results over the issue of educational standards. NAEP data are, or could be, highly pertinent for this purpose. And it puts NAEP in a position, to use Bruner's (1966) words, of providing "the full range of alternatives to challenge society to choice."

IV.
Epilogue

The last chapter of this report of a proposed NAEP redesign focusses on ways to improve NAEP's flexibility for meeting varied assessment needs, with particular stress on heightening NAEP capabilities for

- addressing multiple policy questions,
- reaching multiple audiences in effective fashion,
- linking to other valuable data sources,
- enhancing and extending assessment services, and
- engaging the public around NAEP data on the important social issue of educational standards.

Thus, our closing emphasis is on strategies to improve policy impact, dissemination, knowledge utilization, user services, and public involvement.

But we should not forget that the main reason this closing emphasis is needed was covered in Chapter II. NAEP's perennial difficulties in policy analysis, dissemination, service and knowledge utilization, and public engagement stem directly from the design problems addressed there. The original design led to performance data that lacked direct comparability across exercises, age levels, population subgroups, and time periods as well as to the results of other assessment programs. This resulted in findings of debatable meaning that were difficult to interpret, especially with respect to time trends. It is not surprising that such data have had little impact on American education.

The proposed redesign remedies these problems by means of BIB spiralling and IRT scaling. This makes possible the formation of meaningful scales whose construct validity, and hence interpretability, can be appraised empirically. It also enables the development of scales with common meaning across exercises, age

levels, subgroups, and time periods, thereby permitting powerful comparisons with clear implications.

Furthermore, the proposed redesign—not only of data collection and analysis procedures, but of reporting, dissemination, and utilization procedures—is accomplished in ways that are

- **protective** of the links to past NAEP data,
- **innovative** in its move to new psychometric methodology, and
- **aggressive** in its outreach.

V.
References

Acland, H. An analysis of NAEP. Unpublished paper prepared for the National Academy of Sciences Committee on Ability Testing, 1980.

Adams, E. K. *A changing federalism: The condition of the states* (Report #[illegible]-1). Denver, CO: Education Commission of the States, 1982.

Ap[illegible]on, R. E., Welch, W. W., & Harris, L. J. Methodological considera[illegible]n the development of indicators of achievement in data from [illegible]ional Assessment, *Journal of Educational Measurement*, 1982, [illegible] 113-124.

Ba[illegible], S. K. & Mosher, E. K. *ESEA: The Office of Education administers a Law*. Syracuse, NY: Syracuse University Press, 1968.

Baratz, J. C. Policy implications of minimum competency testing. In R. Jaeger & C. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, consequences*. Berkeley, CA: McCutchan, 1980.

Baratz, J. C., & Hartle, T. W. "Malpractice" in the schools. *The Progressive*, June, 1977, 33-34.

Bayh, B. *Challenge for the third century: Education in a safe environment—Final report on the nature and prevention of school violence and vandalism*. U. S. Senate Judiciary Committee, Subcommittee on Juvenile Delinquency. Washington, DC: U.S. Government Printing Office, 1977.

Beaton, A. E., Hilton, T. L., & Schrader, W. B. *Changes in the verbal abilities of high school seniors, college entrants, and SAT candidates between 1960 and 1972*. New York: College Entrance Examination Board, 1977.

Bentler, P. M. Multivariate analysis with latent variables: Causal modeling. *Annual review of Psychology*, 1980, **31**, 419-456.

Bishop, Y. M., Feinberg, S., & Holland, P. W. *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press, 1975.

Bruner, J. S. *Toward a theory of instruction*. Cambridge, MA: Harvard University Press, 1966.

Bock, R. D., Mislevy, R., & Woodson, C. The next stage in educational assessment *Educational Researcher*, 1982, **11**(3), 4-11, 16.

V. References

Acland, H. An analysis of NAEP. Unpublished paper prepared for the National Academy of Sciences Committee on Ability Testing, 1980.

Adams, E. K. *A changing federalism: The condition of the states* (Report #F82-1). Denver, CO: Education Commission of the States, 1982.

Anderson, R. E., Welch, W. W., & Harris, L. J. Methodological considerations in the development of indicators of achievement in data from the National Assessment, *Journal of Educational Measurement*, 1982, **19**, 113-124.

Bailey, S. K. & Mosher, E. K. *ESEA: The Office of Education administers a law*. Syracuse, NY: Syracuse University Press, 1968.

Baratz, J. C. Policy implications of minimum competency testing. In R. Jaeger & C. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, consequences*. Berkeley, CA: McCutchan, 1980.

Baratz, J. C., & Hartle, T. W. "Malpractice" in the schools. *The Progressive*, June, 1977, 33-34.

Bayh, B. *Challenge for the third century: Education in a safe environment—Final report on the nature and prevention of school violence and vandalism*. U. S. Senate Judiciary Committee, Subcommittee on Juvenile Delinquency. Washington, DC: U.S. Government Printing Office, 1977.

Beaton, A. E., Hilton, T. L., & Schrader, W. B. *Changes in the verbal abilities of high school seniors, college entrants, and SAT candidates between 1960 and 1972*. New York: College Entrance Examination Board, 1977.

Bentler, P. M. Multivariate analysis with latent variables: Causal modeling. *Annual review of Psychology*, 1980, **31**, 419-456.

Bishop, Y. M., Feinberg, S., & Holland, P. W. *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press, 1975.

Bruner, J. S. *Toward a theory of instruction*. Cambridge, MA: Harvard University Press, 1966.

Bock, R. D., Mislevy, R., & Woodson, C. The next stage in educational assessment *Educational Researcher*, 1982, **11**(3), 4-11, 16.

Carroll, J. B. The nature of data, or how to choose a correlation coefficient. *Psychometrika*, 1961, **26**, 347-372.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. *Equality of educational opportunity*. U.S. Office of Education, Department of Health, Education, and Welfare, Washington, DC: U.S. Government Printing Office, 1966.

College Board. *On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline*. New York: The author, 1977.

Cochran, W. C., & Cox, G. M. *Experimental designs* (2nd ed.). New York: Wiley, 1957.

Cronbach, L. J. Test validation. In R. L. Thorndike, *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977, **39**, 1-38.

Dyer, H. S. The discovery and development of educational goals. *Proceedings of the 1966 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1967.

Fiske, E. R. The high schools: New shapes for the eighties. *New York Times*, April 26, 1981.

Frederiksen, N., & Evans, F. R. Effects of models of creative performance on ability to formulate hypotheses. *Journal of Educational Psychology*, 1974, **66**, 67-82.

Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1967.

Hambleton, R. K. *Applications of item response models to NAEP mathematics exercise results*. Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research, 1982.

Heller, K. A., Holtzman, W. H., & Messick, S. *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press, 1982.

Heyns, B., & Hilton, T. L. The cognitive tests for high school and beyond: An assessment. *Sociology of Education*, 1982, **55**, 89-1022.

Hill, P., & Kimbrough, J. *The aggregate effects of federal programs*. Santa Monica, CA: Rand, 1981.

Jöreskog, K. G., & Sörbom, D. *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt, 1979.

Jöreskog, K. G., & Sörbom, D. *LISREL V, estimation of linear structural equation systems by maximum likelihood methods: A program.* Chicago, IL: National Educational Resources, 1981.

Kagan, J., & Kogan, N. Individual variation in cognitive processes. In P. H. Mussen (Ed.), *Carmichael's manual of child psychology* (Vol. 1, 3rd ed.). New York: Wiley, 1970.

Kelley, T. L. Ridge-route norms. *Harvard Educational Review*, 1940, **10**, 309-314.

Knapp, T. R. An application of balanced incomplete block designs to the estimation of test norms. *Educational and Psychological Measurement*, 1968, **28**, 265-272.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports, 1957*, **3**, 635-694. (Monograph Supplement 9)

Lord, F. M. Equating test scores—A maximum likelihood solution. *Psychometrika*, 1955, **20**, 193-200.

Lord, F. M. Estimating norms by item-sampling. *Educational and Psychological Measurement*, 1962, **22**, 259-267.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 1974, **39**, 247-264.

Lord, F. M. Test theory and the public interest. *Testing and the public interest: Proceedings of the 1976 ETS Invitational Conference*. Princeton, NJ: Educational Testing Service, 1976.

Lord, F. M. A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1977, **1**, 95-100.

Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum, 1980. (a)

Lord, F. M. Some how and which for practical tailored testing. In L. J. Th. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates*. New York: Wiley, 1980. (b)

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Maher, C., & Bennett, R. *Planning and evaluating special education services*. New York: Prentice-Hall, in press.

Mayeske, G. W., Wisler, C. E., Beaton, A. E., Weinfeld, F. D., Cohen, W. M., Okada, T., Proshek, J. M., & Tabler, K. *A study of our nation's schools*. Washington, DC: U.S. Office of Education, U.S., Department of Health, Education, and Welfare, U.S. Government Printing Office, 1972.

McDonald, R. P. Exploratory and confirmatory nonlinear common factor analysis. In S. Messick & H. Wainer (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord.* Hillsdale, NJ: Erlbaum, 1983.

MacDonald-Ross, M. Research in graphic communication. *IET Monograph No. 7*, 1978.

McDonnell, L. N., & McLaughlin, M. W. *Education policy and the role of the states.* Santa Monica, CA: Rand, 1982.

McGill, R., Tukey, J. W., & Larsen, W. Variations of box plots. *American Statistician*, 1978, **32**, 12-16.

McLaughlin, M. W. *State involvement in local education quality issues* (Interim report). Santa Monica, CA: Rand, 1981.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 1975*, **30**, 955-966.

Messick, S. Test validity and the ethics of assessment. *American Psychologist, 1980*, **35**, 1012-1027.

Messick, S., & Barrows, T. S. Strategies for research and evaluation in early childhood education. In I. J. Gordon (Ed.), *Early childhood education: The seventy-first yearbook of the National Society for the Study of Education.* Chicago: University of Chicago Press, 1972.

Milrod, M. *Report of NIE grant application review panel Assessment Policy Committee meeting.* Washington, DC: National Institute of Education, Mimeo, February 9, 1980.

Moore, M. T., Walker, L., & Holland, R. *Finetuning special education: A finance guide for state policymakers.* Washington, DC: Education Policy Research Institute, Educational Testing Service, 1982.

Murphy, J. T. Title V of ESEA: The impact of discretionary funds on state education bureaucracies. *Harvard Educational Review*, 1973, **43**, 362-386.

Odden, A., & Dougherty, V. *State programs of school improvement: A 400 state survey* (Report #182-3). Denver, CO: Education Commission of the States, 1982.

Phi Delta Kappan. Fifth annual Gallup Poll of attitudes toward education—1973. In S. M. Elam (Ed.), *A decade of Gallup Polls of attitudes toward education 1969-1978. Bloomington, IN: Phi Delta Kappa, 1978.*

Research Triangle Institute. The National Assessment of Educational Progress: District supervisor's training manual (Year 11). Durham, NC: The author, 1979.

Rock, D. A., Werts, C., & Grandy, J. E. *Construct validity of the GRE Aptitude Test across populations—An empirical confirmatory study* (ETS *RR 81-57*; GREB Report No. 78-1P). Princeton, NJ: Educational Testing Service, 1981.

Samejima, F. A general model for free-response data. *Psychometric Monographs, 1972,* (Whole No. 18).

Samejima, F. Homogeneous case of the continuous response model. *Psychometrika,* 1973, **38**, 203-219.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika,* 1974, **39**, 111-121.

Sebring, P. A., & Boruch, R. F. *On the uses of the National Assessment of Educational Progress* (Report No. A-137-4). Evanston, IL: Division of Methodology and Evaluation Research, Psychology Department, Northwestern University, 1982.

Shulins, N. The states search for fiscal light in gloom of recession. *Washington Post,* July 13, 1982.

Silverstein, R., et al. *A description and analysis of the relationship between Title I and selected state compensatory education programs.* Washington, DC: Lawyers' Committee for Civil Rights Under Law, 1977.

Smith, M. S. Equality of educational opportunity: The basic findings reconsidered. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity.* New York: Vintage Books, 1972.

Smith, R. L. (Council for the Advancement of Private Education). Speech presented to the annual meeting of the American Association of Secondary School Administrators, New Orleans, February 1982.

Stocking, M. L., & Lord, F. M. Developing a common metric in item response theory. *Applied Psychological Measurement,* in press.

Stricker, L. J. Interpersonal competence instrument: Development and preliminary findings. *Applied Psychological Measurement, 1982,* **6**, 69-81.

Sundquist, J., & Davis, D. *Making federalism work.* Washington, DC: The Brookings Institution, 1969.

Tucker, L. R Searching for structure in binary data. In S. Messick & H. Wainer (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord.* Hillsdale, NJ: Erlbaum, 1983.

Tukey, J. W. *Exploratory data analysis.* Reading , MA: Addision-Wesley, 1977.

Ward, W. C. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, 1982, 6, 1-11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. Construct validity of free-response and machine-scorable versions of a test of scientific thinking. *Journal of Educational Measurement*, 1980, 17, 11-29.

Wainer, H., & Thissen, D. Graphical data analysis. *Annual Review of Psychology*, 1981, 32, 191-241.

Wainer, H., Groves, C., & Lono, M. Some experiments in graphical comprehension. Paper presented at the annual meeting of the American Statistical Association, San Diego, 1978.

Wainer, H., Groves, C., & Lono, M. On the display of data: Some empirical findings. Unpublished manuscript, 1979.

Wiley, D. E. Improving policy development. *New Directions for Testing and Measurement: Testing in the states, beyond accountability*, 10. San Francisco: Jossey-Bass, 1981.

Wilken, W. H., & Porter, D. O. *State aid for special education: Who benefits?* Washington, DC: National Institute for Education, 1977.

Wingersky, B. Innovation in multivariate statistics. Paper in preparation, 1982.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test modes. In R. K. Hambleton (Ed.), *ERIBC Monograph on Applications of Item Response Theory*. Vancouver, BC: Educational Research Institution of British Columbia, 1982.

Wingersky, M. S., Barton, M. A., & Lord, F. M. *LOGIST user's guide: LOGIST 5, Verson 1.0*. Princeton, NJ: Educational Testing Service, 1982.

Wirtz, W., & Lapointe, A. *Measuring the quality of education: A report on assessing educational progress*. Washington, DC: The authors, 1982.